ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΤΣΙΩΝΟΣ ΓΕΩΡΓΙΟΣ-ΠΑΝΑΓΙΩΤΗΣ

ΛΑΡΙΣΑ – 2018

Πρόβλεψη θέσεων φωσφορυλίωσης σε πρωτεΐνες του αρουραίου με μεθόδους μηχανικής μάθησης

Prediction of phosphorylation sites in rat proteins with machine learning methods

Πρόβλεψη θέσεων φωσφορυλίωσης σε πρωτεΐνες του αρουραίου με μεθόδους μηχανικής μάθησης

Γεώργιος-Παναγιώτης Τσιώνος

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Γρήγορης Αμούτζιας

Εργαστήριο Βιοπληροφορικής

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ

Γρηγόρης Αμούτζιας (επιβλέπων), Επίκουρος καθηγητής Βιοπληροφορικής στη Γενωμική, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας

Νικόλαος Παπανικολάου, Επίκουρος Καθηγητής Βιολογικής Χημείας, Τμήμα Ιατρικής, Α.Π.Θ.

Ιωάννης Ηλιόπουλος, Επίκουρος Καθηγητής Μοριακής Βιολογίας-Γονιδιωματικής Βιοπληροφορικής, Τμήμα Ιατρικής, Πανεπιστήμιο Κρήτης

Περίληψη

Η φωσφορυλίωση σερίνης, θρεονίνης και τυροσίνης σε πρωτεΐνες είναι μια μετα-μεταφραστική τροποποίηση μεγάλης σημασίας για τα κύτταρα, καθώς μπορεί να επηρεάσει λειτουργίες όπως η ενζυμική δραστικότητα, η δημιουργία συμπλόκων και η μεταγωγή σήματος. Δεδομένου ότι ακόμα δεν έχουν εντοπιστεί οι περισσότερες θέσεις φωσφορυλίωσης στα πρωτεώματα του ανθρώπου και άλλων σημαντικών οργανισμών μοντέλων, όπως π.χ, ο αρουραίος, η πρόβλεψη αυτών των θέσεων είναι πολύ σημαντική για τον σχεδιασμό μελλοντικών πειραμάτων. Στην παρούσα εργασία συλλέξαμε και φιλτράραμε δεδομένα φωσφορυλίωσης του οργανισμού *Rattus norvegicus* από πειράματα φωσφοπρωτεωμικής μεγάλης κλίμακας και τα χρησιμοποιήσαμε για να αξιολογήσουμε διάφορες μεθόδους μηχανικής μάθησης, για την πρόβλεψη της φωσφορυλίωσης πρωτεϊνών του αρουραίου. Από τις διάφορες μεθόδους και εργαλεία που εφαρμόσαμε, τα νευρωνικά δίκτυα που κατασκευάστηκαν στην Matlab, είχαν την μεγαλύτερη ακρίβεια (79,8%). Ωστόσο, και τα νευρωνικά δίκτυα που κατασκευάστηκαν με το Keras/Tensorflow πέτυχαν παρόμοια ακρίβεια (78,6%). Από 7 αλγόριθμους του WEKA, η λογιστική παλινδρόμιση πέτυχε ακρίβεια της τάξης του 78,9%. Από τα παραπάνω συμπεραίνουμε ότι για το συγκεκριμένο πρόβλημα, οι διάφοροι αλγόριθμοι έχουν παρόμοια επιτυχία και ίσως ο πιο καθοριστικός παράγοντας είναι τα δεδομένα εκπαίδευσης και η ποιότητά τους.

Abstract

Protein phosphorylation is a very important post-translational modification, because it may regulate enzyme activity, complex formation or signal transduction. Up to now, the majority of phosphorylation sites in human and other model organisms such as the rat (*Rattus norvegicus*) has not been identified. Therefore, predicting these sites is important for wet-lab experimentalists. We have gathered and filtered published high-throughput phosphoproteomic data from rat and have used them to train various machine-learning algorithms, in order to assess their efficacy to predict phosphorylation sites in this species. Among the various tools that we have utilized, the neural networks that were trained in Matlab achieved the highest accuracy (79.8%). Keras/Tensorflow neural networks also achieved similar levels of accuracy (78.6%). The WEKA software suite was also used to assess 7 different machine–learning algorithms, with logistic regression achieving the highest accuracy among them (78.9%). We conclude that the various algorithms achieve comparable results and also speculate that probably the most important factor is the abundance and quality of the training datasets.

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της πτυχιακής εργασίας μου, Επίκουρο Καθηγητή κ. Αμούτζια Γρηγόριο, για την αμέριστη βοήθεια και καθοδήγησή του, καθώς και για την εμπιστοσύνη που μου έδειξε κατά τη διάρκεια της δουλειάς και παραμονής μου στο εργαστήριο.

Επίσης, είμαι ευγνώμων στον Επίκουρο Καθηγητή κ. Νικόλαο Παπανικολάου και τον Επίκουρο Καθηγητή κ. Ιωάννη Ιλιόπουλο, για τις πολύτιμες υποδείξεις τους και που με τίμησαν με τη συμμετοχή τους στην τριμελή εξεταστική επιτροπή της πτυχιακής εργασίας μου.

Ευχαριστώ θερμά τον συνάδελφο και υποψήφιο διδάκτορα Παναγιώτη Βλασταρίδη για τη βοήθεια και τη στήριξή του. Τέλος, ευχαριστώ τους γονείς μου, για την υποστήριξη μου παρέχουν όλα αυτά τα χρόνια.

Περιεχόμενα

Πίνακας περιεχομένων για τις εικόνες 7
Πίνακας περιεχομένων για τους πίνακες 8
Εισαγωγή 9
1.1 Η βιολογική σημασία της φωσφορυλίωσης 9
1.1.1 Οι προκλήσεις της πρόβλεψης φωσφορυλίωσης 10
1.2 Μηχανική μάθηση 12
1.2.1 Τεχνητά νευρωνικά δίκτυα 12
Υλικά και μέθοδοι 17
2.1 Τα Δεδομένα 17
2.2 Εκπαίδευση τεχνητού νευρωνικού δικτύου στη Matlab 18
2.3 Αλγόριθμοι μηχανικής μάθησης στο Weka 22
2.4 Εκπαίδευση τεχνητού νευρωνικού δικτύου με Keras/Tensorflow. 24
Αποτελέσματα – Συζήτηση 26
3.1 Αποτελέσματα εκπαίδευσης του τεχνητού νευρωνικού δικτύου στο περιβάλλον Matlab. 26
3.2 Αποτελέσματα ταξινόμησης του Weka 32
3.3 Αποτελέσματα εκπαίδευσης του τεχνητού νευρωνικού δικτύου στο Keras/Tensorflow. 38
Συμπεράσματα 50
Βιβλιογραφία 51

Πίνακας περιεχομένων για τις εικόνες

Εικόνα 1.Η διαφορά κλασικού προγραμματισμού και μηχανικής μάθησης.[36] 12
Εικόνα 2.Το Μοντελο ενός νευρώνα μιας εισόδου [37]..... 13
Εικόνα 3. Μοντέλο S νευρώνων[37]..... 14
Εικόνα 4. Μοντέλο S νευρώνων (συντομευμένη μορφή)[37]. 14
Εικόνα 5. Σχηματική Αναπαράσταση της διαδικασίας μάθησης [36]..... 15
Εικόνα 6. Παράδειγμα εφαρμογής αλγορίθμου σύγκλισης με ελάττωση της παραγώγου σε καμπύλη απώλειας.[36]..... 16
Εικόνα 7.Εισαγωγή των δεδομένων στο Matlab..... 20
Εικόνα 8.Η επιλογή δεδομένων εκπαίδευσης, επικύρωσης και δοκιμής. 21
Εικόνα 9. Επιλογή αριθμού κρυφών νευρώνων. 22
Εικόνα 10. Η μορφή εισαγωγής των δεδομένων στο Weka. 23
Εικόνα 11. Η καρτέλα Classify του Weka. 23
Εικόνα 12. Μερικοί από τους διαθέσιμους Classifiers του WEKA..... 24
Εικόνα 13. Πίνακας σύγχυσης για 4 νευρώνες. 26
Εικόνα 14. Η παράσταση ROC (receiver operating characteristic) για 4 νευρώνες..... 27
Εικόνα 15. Παράσταση μείωσης της διεντροπίας για 4 νευρώνες. 28
Εικόνα 16. Πίνακας σύγχυσης για 11 νευρώνες. 29
Εικόνα 17. Η παράσταση ROC για 11 νευρώνες. 30
Εικόνα 18. Η γραφική παράσταση μείωσης της διεντροπίας για 11 νευρώνες. 31
Εικόνα 19. Αποτελέσματα λογιστικής παλινδρόμησης στο WEKA. 32
Εικόνα 20.Αποτελέσματα ταξινόμησης του Naive Bayes στο WEKA. 33
Εικόνα 21. Αποτελέσματα Ανάλυσης του ταξινομητή K-nearest neighbours στο WEKA. 34
Εικόνα 22. Αποτελέσματα ταξινόμησης του αλγορίθμου Reptree στο WEKA. 35
Εικόνα 23. Αποτελέσματα ταξινόμησης του αλγορίθμου J48 στο WEKA. 36
Εικόνα 24. Αποτελέσματα ταξινόμησης του αλγορίθμου DecisionTable στο WEKA. 37
Εικόνα 25. Η σύνοψη του μοντέλου δύο κρυφών επιπέδων χωρίς Dropout. 38
Εικόνα 26. Ο πίνακας σύγχυσης και η ακρίβεια του μοντέλου δύο κρυφών επιπέδων χωρίς Dropout. 38
Εικόνα 27. Οι γραφικές παραστάσεις της απώλειας του μοντέλου δύο κρυφών επιπέδων χωρίς Dropout. 39
Εικόνα 28. Οι γραφικές παραστάσεις της ακρίβειας του μοντέλου δύο κρυφών επιπέδων χωρίς Dropout. 40
Εικόνα 29. Η σύνοψη του μοντέλου του μοντέλου δύο κρυφών επιπέδων με Dropout 0.2 41
Εικόνα 30. Ο πίνακας σύγχυσης και η ακρίβεια του μοντέλου δύο κρυφών επιπέδων με Dropout 0.2. 41
Εικόνα 31. Οι γραφικές παραστάσεις της απώλειας του μοντέλου δύο κρυφών επιπέδων με Dropout 0.2..... 42
Εικόνα 32. Οι γραφικές παραστάσεις της ακρίβειας του μοντέλου δύο κρυφών επιπέδων με Dropout 0.2..... 43

Εικόνα 33. Η σύνοψη του μοντέλου για 11 κρυφούς νευρώνες. 44
Εικόνα 34. Οι γραφικές παραστάσεις της ακρίβειας για 8 κρυφούς νευρώνες και Dropout ίσο με 0. 45
Εικόνα 35. Οι γραφικές παραστάσεις της ακρίβειας για 8 κρυφούς νευρώνες και Dropout ίσο με 0.2. 45
Εικόνα 36. Οι γραφικές παραστάσεις της ακρίβειας για 8 κρυφούς νευρώνες και Dropout ίσο με 0.4. 46
Εικόνα 37. Οι γραφικές παραστάσεις της ακρίβειας για 11 κρυφούς νευρώνες και Dropout ίσο με 0. 46
Εικόνα 38. Οι γραφικές παραστάσεις της ακρίβειας για 11 κρυφούς νευρώνες και Dropout ίσο με 0.2. 47
Εικόνα 39. Οι γραφικές παραστάσεις της ακρίβειας για 11 κρυφούς νευρώνες και Dropout ίσο με 0.4. 47
Εικόνα 40. Οι γραφικές παραστάσεις της ακρίβειας για 14 κρυφούς νευρώνες και Dropout ίσο με 0. 48
Εικόνα 41. Οι γραφικές παραστάσεις της ακρίβειας για 14 κρυφούς νευρώνες και Dropout ίσο με 0.2. 48
Εικόνα 42. Οι γραφικές παραστάσεις της ακρίβειας για 14 κρυφούς νευρώνες και Dropout ίσο με 0.4. 49

Πίνακας περιεχομένων για τους πίνακες

Πίνακας 1. Οι εργασίες από τις οποίες συλλεχτήκαν τα δεδομένα φωσφορυλίωσης, ... 17
Πίνακας 2. Onehot κωδικοποίηση των αμινοξέων 19
Πίνακας 3. Η ακρίβεια πρόβλεψης των θέσεων φωσφορυλίωσης στον αρουραίο για 7 αλγορίθμους ταξινόμησης του WEKA. 37
Πίνακας 4. Η ακρίβεια για κάθε συνδυασμό κρυφών νευρώνων και Dropout. 44

Εισαγωγή

1.1 Η βιολογική σημασία της φωσφορυλίωσης

Την τελευταία δεκαετία με την πρόοδο που έχει επιτευχτεί στην υψηλής απόδοσης φωσφοπρωτεωμική (HTP, high throughput phosphoproteomics), εκατοντάδες ή και χιλιάδες θέσεις φωσφορυλίωσης (phosphorylation sites - p-sites) μπορούν να ανιχνευθούν σε ένα μόνο πείραμα. Η επιτυχία αυτή οφείλεται στο συνδυασμό πολύ ευαίσθητων οργάνων φασματομετρίας μάζας, καλύτερες τεχνικές εμπλουτισμού φωσφοπεπτιδίων και καλύτερου λογισμικού βιοπληροφορικής για την ανάλυση των δεδομένων.

Για την κατανόηση ενός βιολογικού συστήματος δεν αρκεί να ξέρουμε ποια μόρια εκφράζονται σε διαφορές καταστάσεις. Οι πρόσφατες εξελίξεις στην υψηλής απόδοσης πρωτεωμική και φωσφοπρωτεωμική έχουν επισημάνει τη σημασία του να γνωρίζουμε εάν οι εκφρασμένες πρωτεΐνες έχουν τις μοριακές τους λειτουργίες ενεργοποιημένες ή απενεργοποιημένες μέσω μετα-μεταφραστικών τροποποιήσεων. Η φωσφορυλίωση είναι η πιο άφθονη αναστρέψιμη μετα-μεταφραστική τροποποίηση [1]. Μπορεί να λειτουργήσει σαν ψηφιακός διακόπτης ή σαν ρεοστάτης, ρυθμίζοντας μια ή περισσότερες λειτουργίες σε μια πρωτεΐνη, όπως η ενζυμική δραστηριότητα και ο σχηματισμός συμπλόκων. Η φωσφορυλίωση / αποφωσφορυλίωση παίζει επίσης μεγάλο ρόλο στην μεταγωγή σήματος. Μπορεί να υπάρχουν περισσότεροι από ένας διακόπτες αυτού του είδους σε μια πρωτεΐνη. Μπορεί να είναι ανεξάρτητοι ο ένας από τον άλλο, να υπάρχουν αλληλεξαρτήσεις μεταξύ τους ή ακόμη και με άλλους τύπους διακοπτών. Πρόσφατες μελέτες εκτίμησαν [2-4] ότι 1/3 με 2/3 των πρωτεϊνών σε ένα ευκαρυωτικό οργανισμό αναμένεται να φωσφορυλιώνονται. Επιπλέον μια πρωτεΐνη μπορεί να έχει από μια μέχρι δεκάδες θέσεις φωσφορυλίωσης. Η πολυπλοκότητα που δημιουργείται με τη φωσφορυλίωση είναι τεράστια.

Η μετάλλαξη μόνο μιας θέσης φωσφορυλίωσης σε μια πρωτεΐνη μπορεί να έχει δραματικές επιπτώσεις όχι μόνο για τη λειτουργία της συγκεκριμένης πρωτεΐνης, αλλά και για τα μονοπάτια στα οποία εμπλέκεται ή ακόμη και για το φαινότυπο του οργανισμού [5, 6]. Επιπλέον, οι σημειακές μεταλλάξεις των θέσεων φωσφορυλίωσης στα ένζυμα κλειδιά μπορεί να μεταβάλουν τη ροή των βιοχημικών μονοπατιών του κυττάρου προς επιθυμητά βιοτεχνολογικά προϊόντα ή ιδιότητες [7, 8].

Η μη φυσιολογική φωσφορυλίωση των πρωτεϊνών εμπλέκεται σε πολλές ασθένειες, όπως ο καρκίνος, ο διαβήτης, αυτοάνοσα νοσήματα, καρδιαγγειακές και νευροεκφυλιστικές ασθένειες [9,10]. Νέες γενιές φαρμάκων κατά του καρκίνου και άλλων ασθενειών στοχεύουν τις μετα-μεταφραστικές τροποποιήσεις, ενώ υπάρχει έντονο ενδιαφέρον για τη μέτρηση των φωσφοπρωτεϊνών ορού ή αίματος, με σκοπό τη βελτιωμένη διάγνωση.

Έτσι, η φωσφορυλίωση φαίνεται να είναι ένα ελκυστικό πεδίο έρευνας, όχι μόνο για την κατανόηση της πολυπλοκότητας των οργανισμών ή του τρόπου ρύθμισης των κύτταρων, αλλά επίσης για τη θεραπεία ασθενειών ακόμη και για τη συνθετική βιολογία. Η κατανόησή της θα μας επιτρέψει να ελέγξουμε μοριακά μονοπάτια και φαινότυπους μέσω σημειακών μεταλλάξεων που τροποποιούν ένα μικρό αριθμό κρίσιμων θέσεων φωσφορυλίωσης.

1.1.1 Οι προκλήσεις της πρόβλεψης φωσφορυλίωσης

Μία σημαντική πρόκληση σχετίζεται με την ποιότητα των φωσφο-πρωτεωμικών δεδομένων. Όπως συμβαίνει με κάθε νέα high throughput τεχνολογία, τα δεδομένα που παράγονται επηρεάζονται από πειραματικές στρεβλώσεις (bias) και θόρυβο. Οι διάφορες τεχνικές εμπλουτισμού φωσφοπεπτιδίων συλλαμβάνουν μόνο ένα μέρος του πλήρους φωσφοπρωτεώματος ενώ επίσης εισάγουν στρεβλώσεις. Έχει καταδειχθεί [11] [12], μελετώντας ένα σύνολο 12 φωσφο-πρωτεωμικών πειραμάτων πάνω στη ζύμη ότι περισσότερες από τις μισές μοναδικές θέσεις φωσφορυλίωσης ταυτοποιηθήκαν μόνο μια φορά, επισημαίνοντας το προβλημα των πιθανών ψευδώς θετικών ή μη λειτουργικών p-sites στις high throughput βάσεις δεδομένων.

Μια άλλη ανησυχία σχετίζεται με την αυστηρότητα των κριτηρίων και των αλγορίθμων που χρησιμοποιούνται για την ταυτοποίηση των φωσφοπεπτιδίων και τον σωστό εντοπισμό των θέσεων φωσφορυλίωσης σε ένα φωσφοπεπτίδιο. Ορισμένες βάσεις δεδομένων, αναλύσεις βιοπληροφορικής ή ακόμη και εργαλεία πρόβλεψης εξάγουν τις θέσεις φωσφορυλίωσης από συμπληρωματικό υλικό δημοσιεύσεων χωρίς εφαρμογή πολύ αυστηρών κριτηρίων. Τέτοιες αναλύσεις βασίζονται συχνά στα κριτήρια που καθορίζονται από μεμονωμένες δημοσιεύσεις, τα οποία δεν είναι ομοιόμορφα.

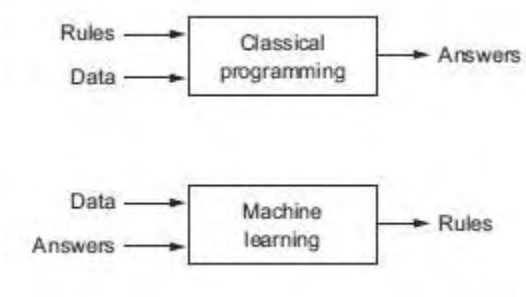
Η αύξηση των HTP φωσφοπρωτεωμικών δεδομένων έδωσε ώθηση στην ανάπτυξη εργαλείων βιοπληροφορικής για την πρόβλεψη θέσεων φωσφορυλίωσης, είτε από την αλληλουχία αμινοξέων μόνο, είτε σε συνδυασμό με δομικές και άλλες πληροφορίες [13-15, 16]. Περισσότερες από 40 μέθοδοι πρόβλεψης έχουν δημοσιευθεί για αυτό το υπολογιστικό πρόβλημα, όπως η εφαρμογή τεχνητών νευρωνικών δικτύων, Support Vector machines, δέντρων αποφάσεων και γενετικών αλγόριθμων. Επιπλέον, υπάρχει πληθώρα βάσεων δεδομένων. Υπάρχουν ακόμα συζητήσεις για το ποιο είναι το βέλτιστο μέγεθος της αλληλουχίας γύρω από το p-site που περιέχει αρκετές πληροφορίες για την κατασκευή ενός μοντέλου πρόβλεψης. Ένα κρίσιμο ζήτημα είναι τα δεδομένα εκπαίδευσης που χρησιμοποιούνται για αυτούς τους αλγορίθμους. Άφθονα και υψηλής ποιότητας p-sites, καθώς και πολύ καλά αρνητικά σύνολα δεδομένων χρειάζονται για την επιτυχή εκπαίδευση αλγορίθμων πρόβλεψης. Δεν αποτελεί έκπληξη το γεγονός ότι τα αρνητικά σύνολα δεδομένων είναι πολύ δύσκολο να ληφθούν από τη στιγμή που ένα μεγάλο μέρος του φωσφοπρωτεώματος ενός οργανισμού

παραμένει άγνωστο. Επίσης, απαιτούνται πολύ καλά σύνολα δεδομένων αναφοράς (τόσο θετικά όσο και αρνητικά) ώστε η κοινότητα να αξιολογήσει κάθε νέο αλγόριθμο / εργαλείο και να το συγκρίνει με τα υπάρχοντα. Μέχρι στιγμής, τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση τέτοιων αλγορίθμων πάσχουν από θορυβώδη p-sites και κακό φιλτράρισμα του τεχνικού και βιολογικού θορύβου. Επιπλέον, τα περισσότερα από τα p-sites που έχουν ληφθεί από τεχνικές high throughput έχουν αναγνωριστεί με πρωτόκολλο που περιλαμβάνει μόνο τη θρυψίνη. Η χρήση δύο ή περισσοτέρων πρωτεολυτικών ενζύμων διαδοχικά (π.χ. Lys-N και θρυψίνη) είχε ως αποτέλεσμα δεδομένα εμπλουτισμένα με περισσότερα μοτίβα φωσφορυλίωσης [17]. Ως εκ τούτου, καθώς χρησιμοποιούνται πιο σύνθετα πρωτόκολλα με περισσότερα από ένα ένζυμα πέψης, περισσότερα μοτίβα φωσφορυλίωσης θα ληφθούν και θα χρησιμοποιηθούν για την εκπαίδευση νέων αλγορίθμων πρόβλεψης.

Παρά τον θόρυβο που υπάρχει στα τρέχοντα σύνολα φωσφο-πρωτεωμικών δεδομένων και τις ελλείψεις τους, τα συμπεράσματα των υπολογιστικών αναλύσεων που έγιναν μέχρι τώρα με περιορισμένα σύνολα δεδομένων δεν είναι αναγκαστικά άκυρα. Όπως καταδεικνύεται στο [2], αρκετά συμπεράσματα παραμένουν ισχυρά, όταν έχουμε μεγάλα και υψηλής ποιότητας σύνολα δεδομένων των p-sites. Παρ 'όλα αυτά, το αποτελεσματικό φιλτράρισμα θορύβου θα αυξήσει σημαντικά την εμπιστοσύνη στα συμπεράσματα σχετικών υπολογιστικών αναλύσεων και έτσι θα επιτρέψει καινούριες ανακαλύψεις.

1.2 Μηχανική μάθηση

Στον κλασικό προγραμματισμό άνθρωποι εισάγουν κανόνες και δεδομένα προς επεξεργασία με σκοπό την εξαγωγή απαντήσεων. Με την μηχανική μάθηση, οι άνθρωποι εισάγουν τα δεδομένα και τις απαντήσεις με σκοπό την εκπαίδευση αλγορίθμων και την εξαγωγή των κανόνων που μετέπειτα θα μπορούν να εφαρμοστούν σε νέα δεδομένα.



Εικόνα 1.Η διαφορά κλασικού προγραμματισμού και μηχανικής μάθησης.[36]

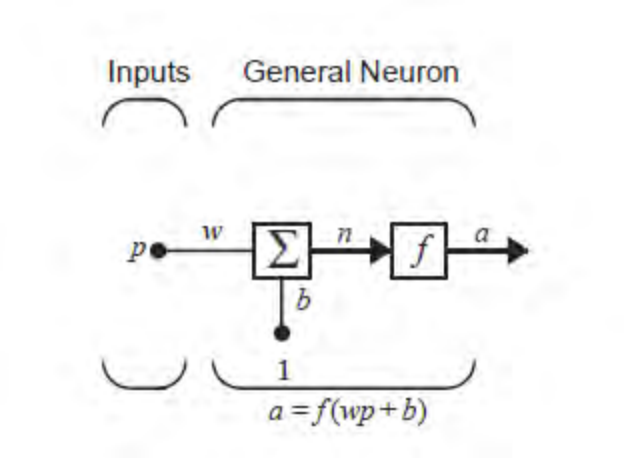
Ένας αλγόριθμος μηχανικής μάθησης χρειάζεται:

- Δεδομένα εισόδου
- Παραδείγματα αναμενόμενων δεδομένων εξόδου
- Έναν τρόπο μέτρησης του σφάλματος ανάμεσα στην έξοδο του αλγορίθμου και την αναμενόμενη έξοδο ώστε να γίνουν οι απαραίτητες διορθώσεις. Αυτή η διαδικασία διόρθωσης ονομάζεται μάθηση.

Σκοπός ενός αλγόριθμου μηχανικής μάθησης είναι να δημιουργήσει νέες αναπαραστάσεις των δεδομένων εισόδου ώστε η έξοδος του να προσεγγίζει τελικά την αναμενόμενη.

1.2.1 Τεχνητά νευρωνικά δίκτυα

Τα τεχνητά νευρωνικά δίκτυα είναι ένα από τα πιο συχνά χρησιμοποιούμενα υπολογιστικά μοντέλα μηχανικής μάθησης. Παρακάτω φαίνεται το μοντέλο ενός νευρώνα μιας εισόδου.



Εικόνα 2.Το Μοντελο ενός νευρώνα μιας εισόδου [37].

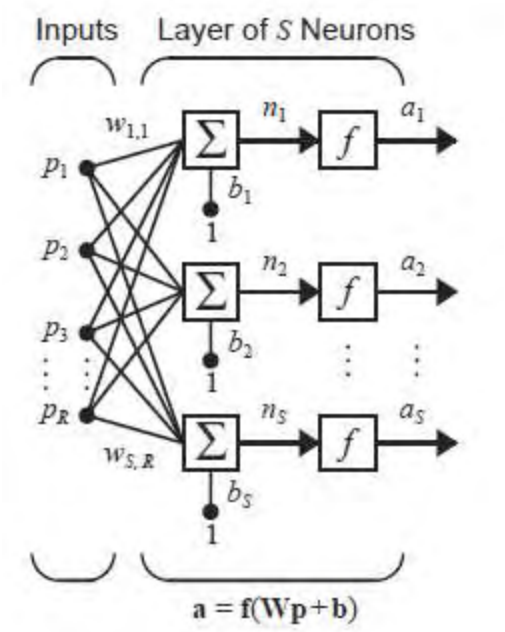
Η είσοδος p πολλαπλασιάζεται με το βάρος w και σχηματίζεται το γινόμενο p το οποίο οδηγείται στον αθροιστή. Η άλλη είσοδος η οποία έχει πάντα τιμή ίση με 1 πολλαπλασιάζεται με το βαθμωτό διάνυσμα πόλωσης b και το γινόμενο τους αποτελεί την άλλη είσοδο του αθροιστή. Η έξοδος του αθροιστή οδηγείται στην συνάρτηση ενεργοποίησης η οποία παράγει την έξοδο του νευρώνα α.

Η έξοδος του νευρώνα υπολογίζεται ως:

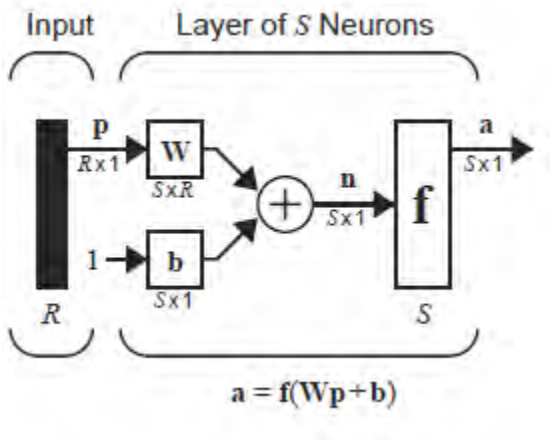
$\alpha = f(wp + b)$.

Το βάρος και το διάνυσμα πόλωσης είναι οι ρυθμιζόμενες παράμετροι του νευρώνα. Οι παράμετροι αυτοί θα ρυθμιστούν κατάλληλα με βάση κάποιο κανόνα μάθησης (Learning rule) ώστε η σχέση εισόδου/εξόδου να ικανοποιεί κάποιο στόχο. Η συνάρτηση ενεργοποίησης ορίζεται από εμάς. Χωρίς αυτή το ΤΝΔ θα μπορούσε να μάθει μόνο γραμμικούς μετασχηματισμούς των δεδομένων εισόδου. Η μη γραμμικότητητα που προσδίδει η χρήση της στο ΤΝΔ, επεκτείνει το χώρο υποθέσεων του ΤΝΔ. Συνήθεις συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στα ΤΝΔ είναι η σιγμοειδής και η συνάρτηση μονάδων γραμμικής ανόρθωσης (relu).

Παρακάτω φαίνεται το μοντέλο ενός επίπεδου των S νευρώνων.

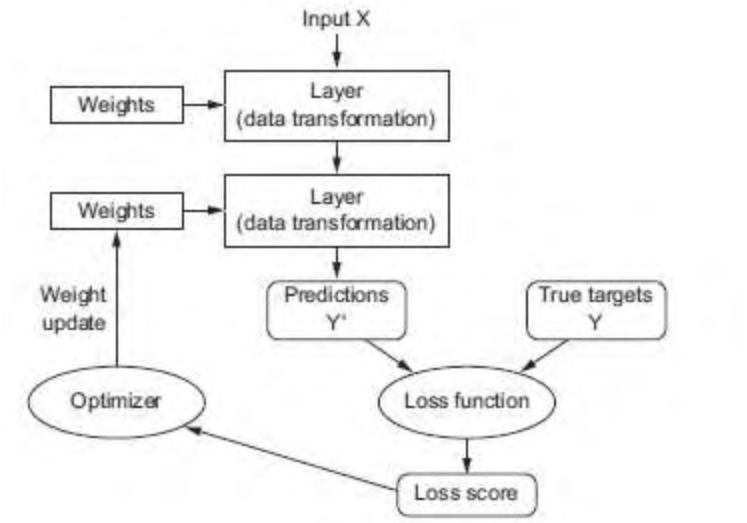


Εικόνα 3. Μοντέλο S νευρώνων[37].



Εικόνα 4. Μοντέλο S νευρώνων (συντομευμένη μορφή)[37].

Εδώ, κάθε μια από τις R εισόδους συνδέεται με καθέναν από τους S νευρώνες. Το επίπεδο αποτελείται από τον πίνακα βαρών, τους αθροιστές, το διάνυσμα πόλωσης τις συναρτήσεις ενεργοποίησης και το διάνυσμα εξόδου α. Κάθε στοιχείο του διανύσματος εισόδου συνδέεται με κάθε νευρώνα μέσω του πίνακα W. Ο κάθε νευρώνας έχει το δικό του διάνυσμα πόλωσης, αθροιστή, συνάρτηση εξόδου και έξοδο a_i ώστε οι έξοδοι όλων των νευρώνων σχηματίζουν το διάνυσμα εξόδου a. Το διάνυσμα p είναι μήκους R, ο πίνακας βαρών διαστάσεων SxR και τα διανύσματα a και b μήκους S. Πολλά τέτοια επίπεδα μπορούν να συνδεθούν διαδοχικά το ένα πίσω από το άλλο ώστε να δημιουργήσουν διάφορες αρχιτεκτονικές ΤΝΔ.

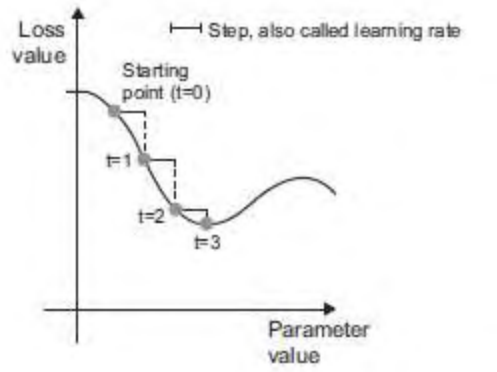


Εικόνα 5. Σχηματική Αναπαράσταση της διαδικασίας μάθησης [36]

Η εκπαίδευση του ΤΝΔ απαιτεί ένα μέτρο αξιολόγησης της επίδοσής του. Το μέτρο αυτό είναι η συνάρτηση απώλειας η οποία υπολογίζει την απόσταση ανάμεσα στις προβλέψεις του ΤΝΔ και τις αναμενόμενες τιμές. Έπειτα, η απόσταση αυτή θα χρησιμοποιηθεί ώστε να ρυθμιστούν οι τιμές των βαρών κατά τέτοιο τρόπο ώστε να μειωθεί η απώλεια. Η ρύθμιση των βαρών υλοποιείται από έναν αλγόριθμο αντίστροφης διάδοσης (Back Propagation). Ο αλγόριθμος αυτός βασίζεται στον υπολογισμό της κλίσης της συνάρτησης απώλειας ώστε να γίνει η διόρθωση των βαρών σε κατεύθυνση αντίθετη της κλίσης. Είναι ένας αλγόριθμος σύγκλισης με ελάττωση της παραγώγου (gradient descent).

Συνοπτικά, στην απλή του μορφή θα μπορούσε να περιγραφεί ως ακολούθως:

- Εισάγουμε μια ομάδα δεδομένων εισόδου και τις αναμενόμενες εξόδους
- Με βάση τα δεδομένα εισόδου κάνουμε προβλέψεις.
- Υπολογίζουμε την απώλεια μεταξύ των αναμενόμενων εξόδων και των προβλέψεων.
- Υπολογίζουμε την κλίση της απώλειας .
- Ανανεώνουμε τις τιμές των βαρών ώστε να μειωθεί η απώλεια [36].



Εικόνα 6. Παράδειγμα εφαρμογής αλγορίθμου σύγκλισης με ελάττωση της παραγώγου σε καμπύλη απώλειας.[36]

Σκοπός αυτής της εργασίας είναι να αξιολογήσουμε διάφορους αλγόριθμους μηχανικής μάθησης για την πρόβλεψη θέσεων φωσφορυλίωσης στον αρουραίο (*Rattus norvegicus*) χρησιμοποιώντας υψηλής ποιότητας δεδομένα φωσφοπρωτεωμικής για εκπαίδευση. Ο αρουραίος επιλέχθηκε γιατί είναι ένας σημαντικός οργανισμός μοντέλο για να κατανοήσουμε την βιολογία του ανθρώπου και των θηλαστικών. Επιπλέον, υπήρχαν αρκετά δεδομένα υψηλής ποιότητας στην βιβλιογραφία.

Υλικά και μέθοδοι

2.1 Τα Δεδομένα

Αρχικά αντλήσαμε δεδομένα φωσφορυλίωσης για τον *Rattus norvegicus* από τις παρακάτω εργασίες: (Δες πίνακα 1.)

Πίνακας 1. Οι εργασίες από τις οποίες συλλεχτήκαν τα δεδομένα φωσφορυλίωσης.

Pubmed ID	Τίτλος
16396499	Phosphoproteomic analysis of rat liver by high capacity IMAC and LC-MS/MS
17683130	An automated platform for analysis of phosphoproteomic datasets: application to kidney collecting duct phosphoproteins
20028136	Phosphoproteome analysis of rat L6 myotubes using reversed-phase C18 prefractionation and titanium dioxide enrichment
20568813	Organic-inorganic hybrid silica monolith based immobilized titanium ion affinity chromatography column for analysis of mitochondrial phosphoproteome
20628157	Signaling to transcription networks in the neuronal retrograde injury response
21630457	ProteoConnections: a bioinformatics platform to facilitate proteome and phosphoproteome analyses
21738781	Phosphoproteomic profiling of in vivo signaling in liver by the mammalian target of rapamycin complex 1 (mTORC1)
22276854	Comprehensive phosphoproteome analysis of INS-1 pancreatic β-cells using various digestion strategies coupled with liquid chromatography-tandem mass spectrometry
22345510	Characterization of membrane-shed microvesicles from cytokine-stimulated β-cells using proteomics strategies
22609512	Novel tyrosine phosphorylation sites in rat skeletal muscle revealed by phosphopeptide enrichment and HPLC-ESI-MS/MS
22673903	Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues
23800682	Testicular phosphoproteome in perfluorododecanoic acid-exposed rats
23984901	Delayed times to tissue fixation result in unpredictable global phosphoproteome changes
24214862	Combined phospho- and glycoproteome enrichment in nephrocalcinosis tissues of phytate-fed rats
24467267	Lifelong exercise training modulates cardiac mitochondrial phosphoproteome in rats
24945867	Salt-induced changes in cardiac phosphoproteome in a rat model of chronic renal failure
25403869	Phosphoproteomic analysis reveals compensatory effects in the piriform cortex of VX nerve agent exposed rats

Τα δεδομένα αυτά φιλτραρίστηκαν ώστε να απομείνουν φωσφοπεπτίδια υψηλής ποιότητας, με ακρίβεια 99% και πάνω. Με τη βοήθεια PERL scripts, αυτά τα φωσφοπεπτίδια στοιχήθηκαν πάνω στο πρωτέωμα του ποντικού που ανακτήθηκε από την βάση δεδομένων ENSEMBL και έτσι εντοπίστηκαν οι θέσεις φωσφορυλίωσης.

Έπειτα, για κάθε μία από τις 12544 θέσεις φωσφορυλίωσης που βρήκαμε δημιουργήσαμε το αντίστοιχο μοτίβο 11 αμινοξέων, με 5 αμινοξέα αριστερά και 5 δεξιά από τη θέση φωσφορυλίωσης γύρω από κάθε σερίνη θρεονίνη ή τυροσίνη. Αυτά αποτέλεσαν το θετικό σύνολο δεδομένων. Επίσης, επιλέξαμε τυχαία τον ίδιο αριθμό σερινών, θρεονινών και τυροσινών που δεν βρέθηκαν να φωσφορυλιώνεται στα πειράματα που συλλέξαμε και εντοπίσαμε τα αντίστοιχα μοτίβα 11 αμινοξέων. Αυτά αποτέλεσαν το αρνητικό σύνολο δεδομένων.

2.2 Εκπαίδευση τεχνητού νευρωνικού δικτύου στη Matlab

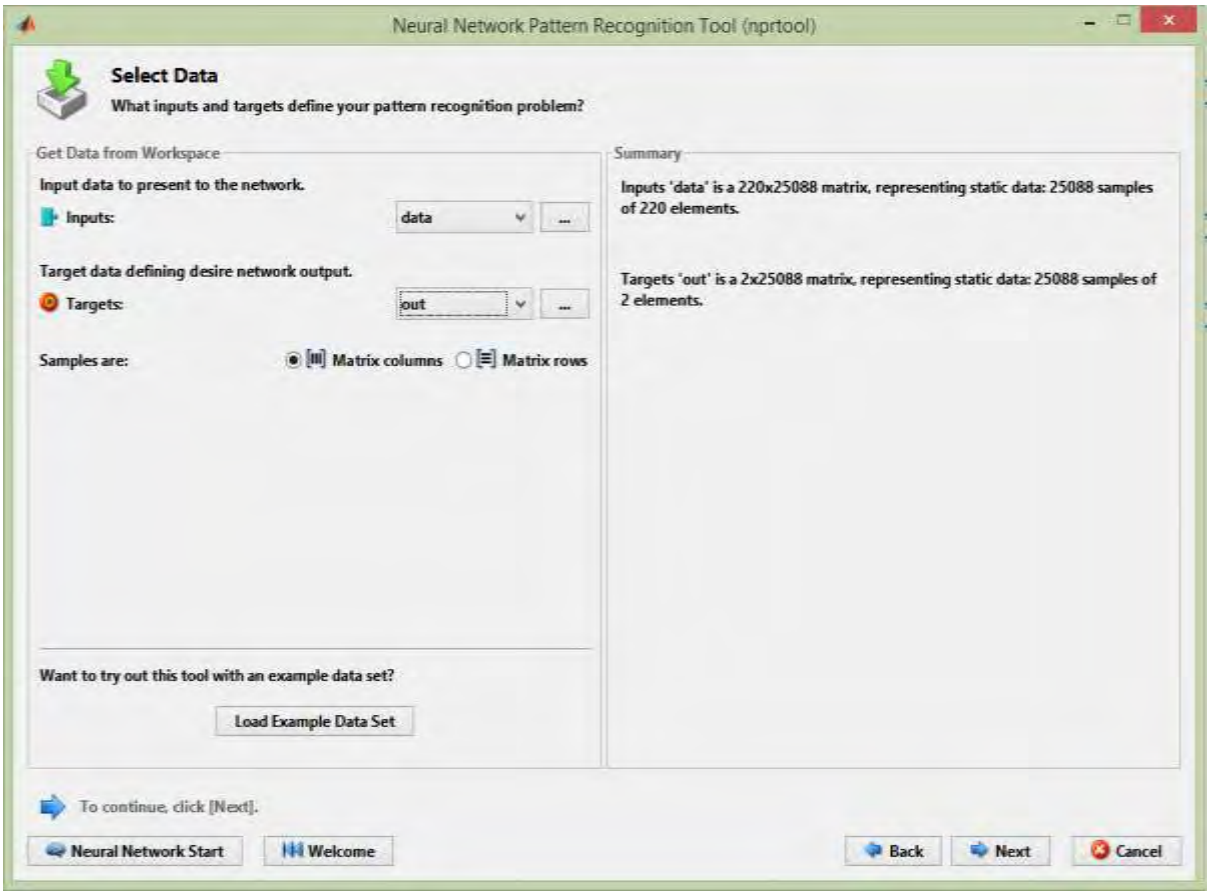
Για την εκπαίδευση του νευρωνικού δικτύου χρησιμοποιήσαμε το εργαλείο Neural Net Pattern Recognition από το Neural Network Toolbox 8.1 του Matlab. Το εργαλείο αυτό κατασκευάζει ένα νευρωνικό δίκτυο που αποτελείται από επίπεδα πρόσω τροφοδότησης, ένα κρυφό επίπεδο με σιγμοειδή συνάρτηση ενεργοποίησης και ένα επίπεδο εξόδου με συνάρτηση ενεργοποίησης τη softmax.

Για να είναι δυνατή η επεξεργασία των δεδομένων μας από το Matlab χρησιμοποιήσαμε την κωδικοποίηση one-hot encoding. Αντιστοιχίσαμε μοναδικά δηλαδή καθένα από τα 20 αμινοξέα στα δεδομένα μας με ένα διάνυσμα 20 στοιχείων όλων μηδενικών εκτός από ένα με τιμή 1. Καταλήξαμε έτσι από μια αλληλουχία 11 αμινοξέων σε ένα διάνυσμα 220 στοιχείων. Για το σύνολο των διαθέσιμων ακολουθιών δημιουργήσαμε έναν πίνακα 220x25088 στοιχείων.

Πίνακας 2. Onehot κωδικοποίηση των αμινοξέων

_	00000000000000000000
A	10000000000000000000
R	01000000000000000000
N	00100000000000000000
D	00010000000000000000
C	00001000000000000000
Q	00000100000000000000
E	00000010000000000000
G	00000001000000000000
H	00000000100000000000
I	00000000010000000000
L	00000000001000000000
K	00000000000100000000
M	00000000000010000000
F	00000000000001000000
P	00000000000000100000
S	00000000000000010000
T	00000000000000001000
W	00000000000000000100
Y	00000000000000000010
V	00000000000000000001

Δημιουργήσαμε επίσης και έναν πίνακα καταστάσεων εξόδου κωδικοποιώντας με 01 την κατάσταση στην οποία δεν έχουμε φωσφορυλίωση και με 10 την κατάσταση στην οποία η ακολουθία φωσφορυλιώνεται. Έπειτα εισάγαμε τα δεδομένα μας στο Matlab όπως φαίνεται στην παρακάτω εικόνα.



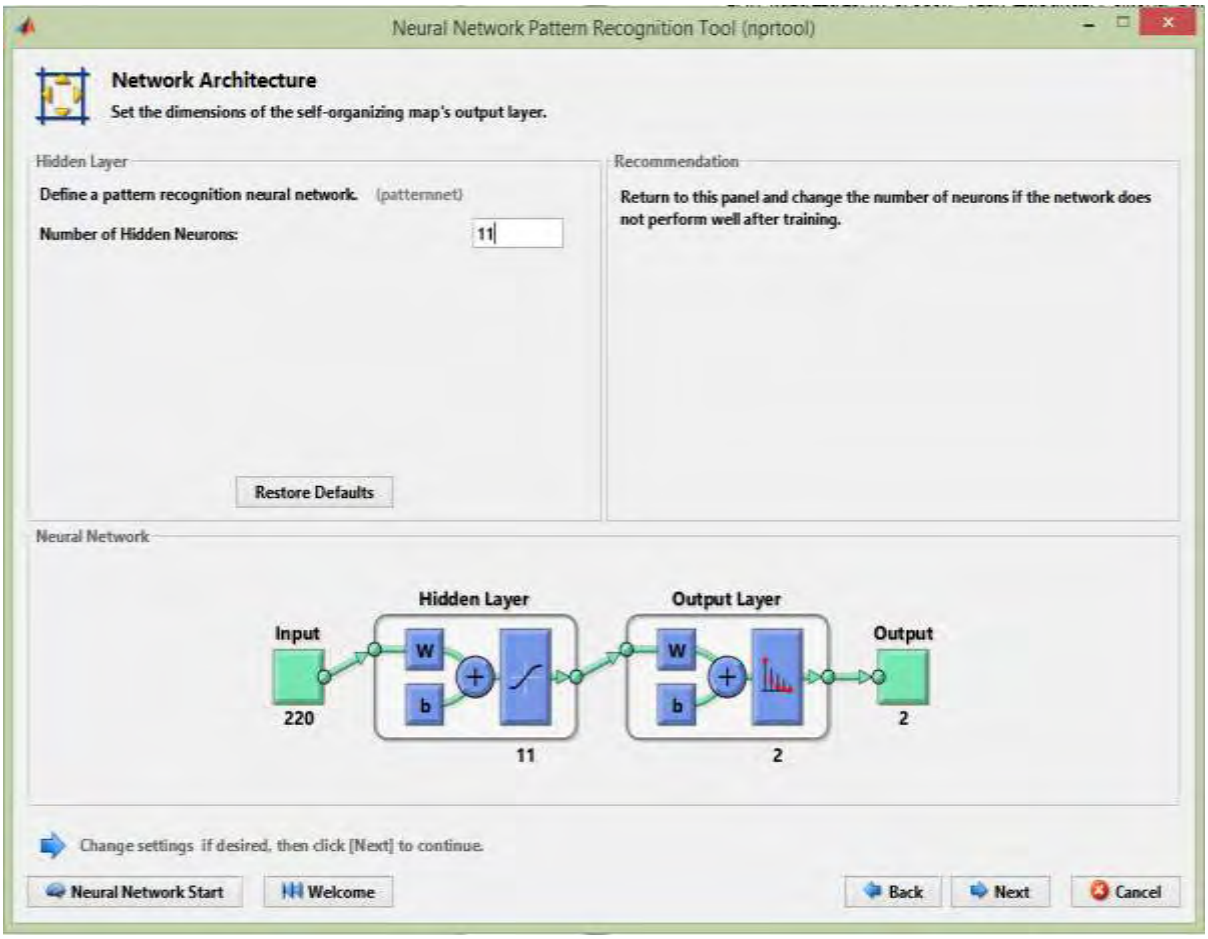
Εικόνα 7.Εισαγωγή των δεδομένων στο Matlab.

Ως inputs εισάγαμε τον πίνακα 220x25088 με τις κωδικοποιημένες αλληλουχίες αμινοξέων και ως targets έναν πίνακα 2x25088 με κωδικοποιημένες τις 2 δυνατές καταστάσεις εξόδου. Έπειτα επιλέξαμε πως θα χωριστούν τα δεδομένα μας για εκπαίδευση (training – 70%), επικύρωση (validation – 15%) και δοκιμή (testing – 15%).

Εικόνα 8.Η επιλογή δεδομένων εκπαίδευσης, επικύρωσης και δοκιμής.

Το νευρωνικό δίκτυο εκπαιδεύεται με τα δεδομένα εκπαίδευσης προσπαθώντας να μειώσει το σφάλμα ανάμεσα σε μια προβλεπόμενη κατάσταση εξόδου και την επιθυμητή κατάσταση εξόδου. Τα δεδομένα επικύρωσης χρησιμοποιούνται ώστε να ληφθεί η απόφαση για τη διακοπή της διαδικασίας εκπαίδευσης πριν το ΤΝΔ να χάσει την ικανότητα γενίκευσης. Τέλος τα δεδομένα εξόδου είναι το τελικό μέτρο εκτίμησης της απόδοσης του νευρωνικού δικτύου.

Το τελευταίο βήμα πριν την εκπαίδευση του νευρωνικού δικτύου είναι να επιλέξουμε τον αριθμό των νευρώνων του κρυφού επιπέδου. Σκοπός μας είναι να επιλέξουμε το βέλτιστο πλήθος νευρώνων ώστε το ΤΝΔ να επιτύχει την καλύτερη δυνατή ακρίβεια στην πρόβλεψη των καταστάσεων εξόδου. Στην παρακάτω εικόνα φαίνεται το παράθυρο επιλογής του αριθμού των κρυφών νευρώνων.



Εικόνα 9. Επιλογή αριθμού κρυφών νευρώνων.

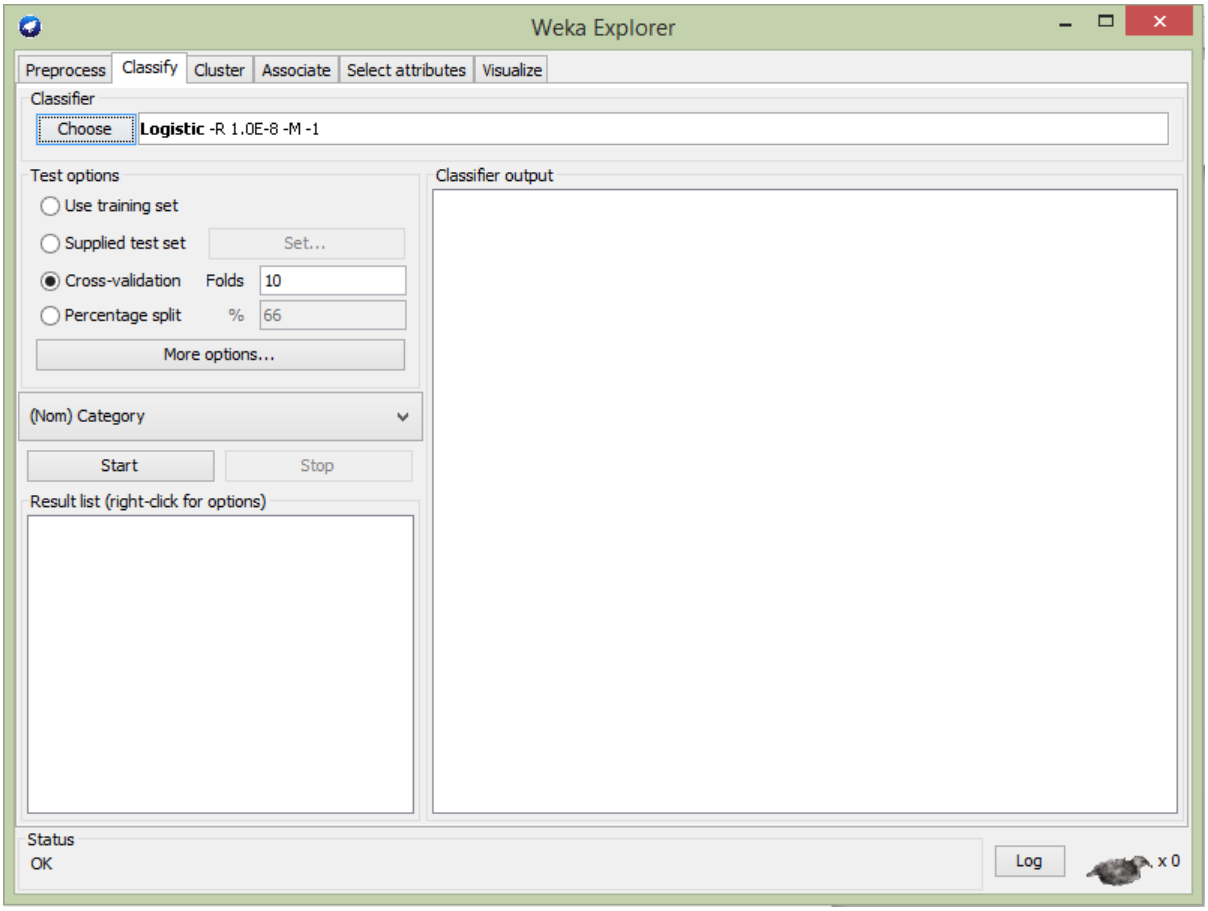
2.3 Αλγόριθμοι μηχανικής μάθησης στο Weka

Το Weka [38] είναι μια εφαρμογή με υλοποιημένους πολλούς αλγόριθμους μηχανικής μάθησης. Στο περιβάλλον του Weka έχουμε τη δυνατότητα αφού εισάγουμε τα δεδομένα μας να επιλέξουμε τον αλγόριθμο που θα χρησιμοποιήσουμε για μηχανική μάθηση. Αρχικά, χρησιμοποιώντας το ίδιο σύνολο δεδομένων με αυτό του Matlab εισάγαμε τα δεδομένα μας σε αρχείο csv της μορφής που φαίνεται στην παρακάτω εικόνα.

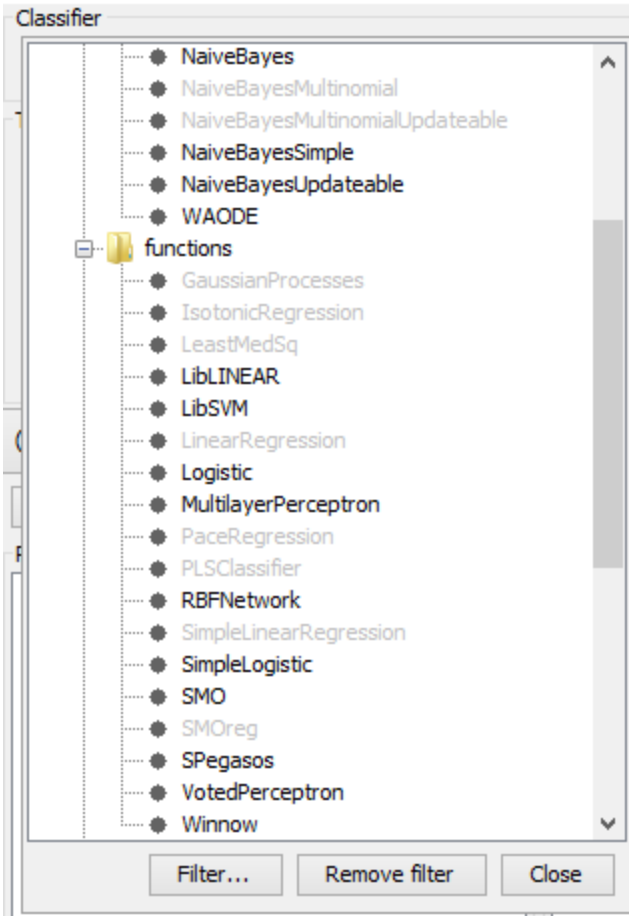
	A	B	C	D	E	F	G	H	I	J	K	L
1	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	Category
2	S	S	S	D	D	S	S	D	E	D	K	POS
3	S	S	D	D	S	S	D	E	D	K	L	POS
4	E	S	D	E	D	S	D	S	D	Q	P	POS
5	M	H	R	A	P	S	P	T	A	E	Q	POS
6	G	S	G	R	E	T	P	Q	P	V	P	POS
7	K	L	N	F	N	S	E	G	E	A	E	POS
8	H	Q	G	K	K	S	I	P	H	I	T	POS
9	V	R	P	V	Q	S	L	P	D	V	C	POS
10	L	T	H	S	S	S	G	N	S	L	K	POS
11	S	S	S	G	N	S	L	K	R	P	D	POS

Εικόνα 10. Η μορφή εισαγωγής των δεδομένων στο Weka.

Έπειτα, επιλέγοντας από το περιβάλλον του Weka την καρτέλα Classify και επιλέγοντας το κουμπί Choose ορίσαμε τους αλγορίθμους ταξινόμησης. Στην καρτέλα Classify έχουμε ακόμη τη δυνατότητα να ορίσουμε το πώς θα γίνει η διαδικασία της επικύρωσης (validation). Στα πειράματα μας επιλέξαμε να χρησιμοποιήσουμε τους αλγορίθμους λογιστικής παλινδρόμησης (Logistic), NaiveBayes, K-nearest neighbor RepTree, J48 και Decision Table.



Εικόνα 11. Η καρτέλα Classify του Weka.



Εικόνα 12. Μερικοί από τους διαθέσιμους Classifiers του WEKA.

2.4 Εκπαίδευση τεχνητού νευρωνικού δικτύου με Keras/Tensorflow.

Το Keras είναι μια βιβλιοθήκη για την ανάπτυξη μοντέλων νευρωνικών δικτύων. Το Tensorflow διαχειρίζεται τις χαμηλού επιπέδου πράξεις μεταξύ τανυστών και χρησιμοποιείται από το Keras σαν backend engine. Το Keras εγκαθίσταται και λειτουργεί σε περιβάλλον Python αφού έχει γίνει εγκατάσταση του Tensorflow.

Χρησιμοποιήσαμε το ίδιο σετ δεδομένων με αυτό του Matlab, δηλαδή 25088 θετικές και αρνητικές για φωσφορυλίωση ακολουθίες, με κωδικοποίηση one-hot. Αυτή τη φορά κρατήσαμε το 80% των δεδομένων μας για εκπαίδευση και επικύρωση και το 20% για δοκιμή.

Για τη δημιουργία ενός μοντέλου στο Keras αποτελούμενου από πλήρως συνδεδεμένα (Dense) επίπεδα αρχικά κάναμε την εισαγωγή των απαραίτητων modules στην Python με τις εντολές.

- `from keras.models import Sequential`
- `from keras.layers import Dense`

΄Επειτα, δημιουργήσαμε διάφορες αρχιτεκτονικές τεχνητών νευρωνικών δικτύων με το επίπεδο Dense το οποίο δημιουργείται με κλήση εντολής του τύπου

- `keras.layers.Dense(units,activation=None,use_bias=True,kernel_initializer='glorot_uniform',bias_initializer='zeros',kernel_regularizer=None,bias_regularizer=None, activity_regularizer=None, kernel_constraint=None, bias_constraint=None)`

Μετά τη δημιουργία του μοντέλου πρέπει να κάνουμε compile ορίζοντας τη συνάρτηση απώλειας και τον βελτιστοποιητή (optimizer).Η συνάρτηση απώλειας καθορίζει το μέγεθος που θα προσπαθήσει το μοντέλο μας να ελαχιστοποιήσει ενώ ο βελτιστοποιητής είναι ο αλγόριθμος αντίστροφης τροφοδότησης που θα ρυθμίζει σε κάθε επανάληψη τις τιμές των βαρών του δικτύου ώστε να ελαχιστοποιηθεί η συνάρτηση απώλειας. Με την παράμετρο metrics ορίζουμε το μέγεθος που θέλουμε να παρακολουθεί το μοντέλο μας κατά την εκπαίδευση. Το μέγεθος αυτό είναι συνήθως η ακρίβεια (accuracy). Η διαδικασία του compile γίνεται με εντολή του παρακάτω τύπου.

- `compile(self, optimizer, loss=None, metrics=None, loss_weights=None, sample_weight_mode=None, weighted_metrics=None, target_tensors=None)`

Το τελευταίο βήμα για την εκπαίδευση του ΤΝΔ είναι η εντολή fit στην όποια καθορίζουμε τα δεδομένα εκπαίδευσης, τον αριθμό των εποχών, το μέγεθος της ομάδας δεδομένων που θα εκπαιδεύεται κάθε φορά και τα δεδομένα που θα χρησιμοποιηθούν για επικύρωση.

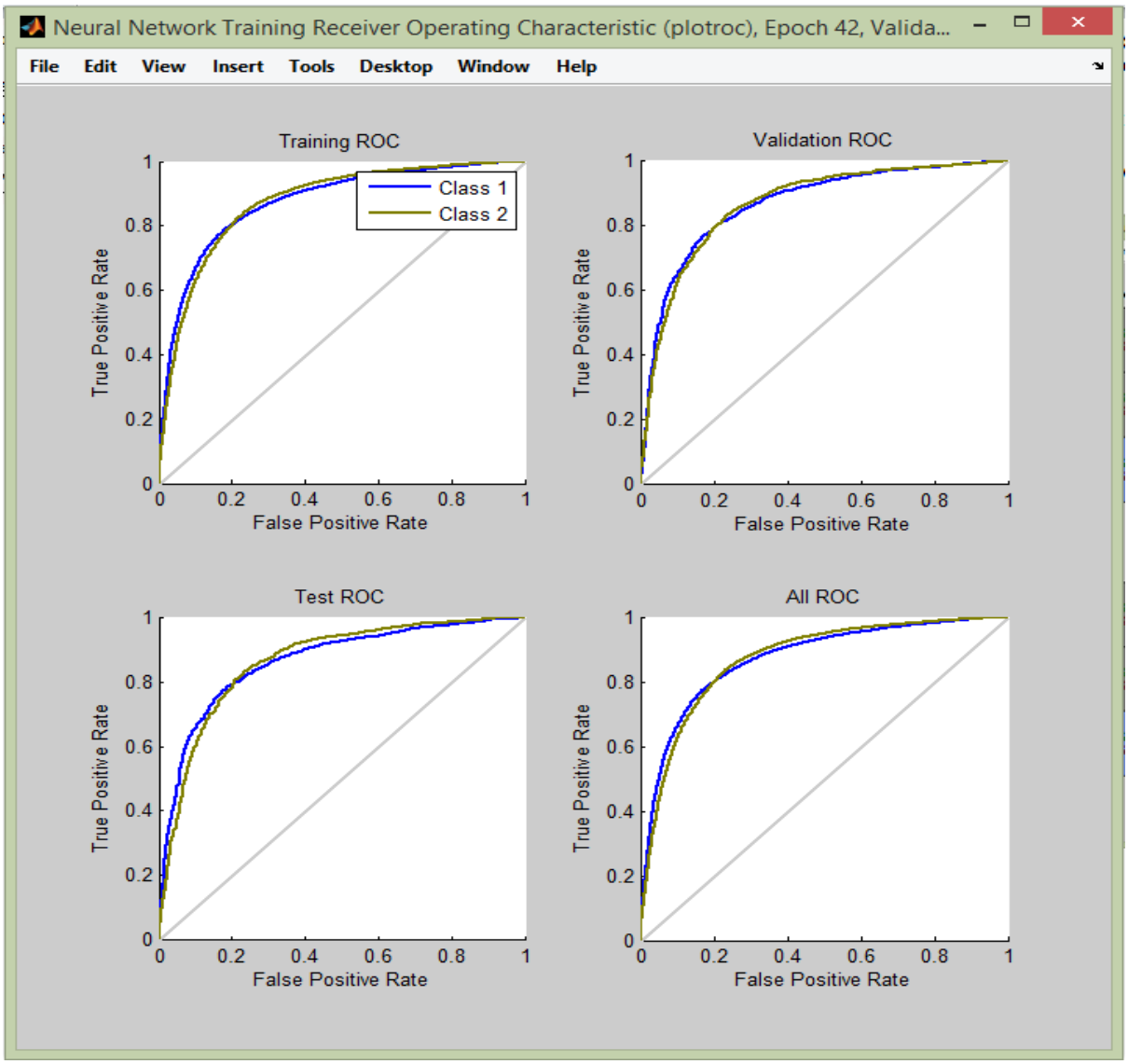
- `fit(self, x=None, y=None, batch_size=None, epochs=1, verbose=1, callbacks=None, validation_split=0.0, validation_data=None, shuffle=True, class_weight=None, sample_weight=None,initial_epoch=0,steps_per_epoch=None, validation_steps=None)`

Αποτελέσματα – Συζήτηση

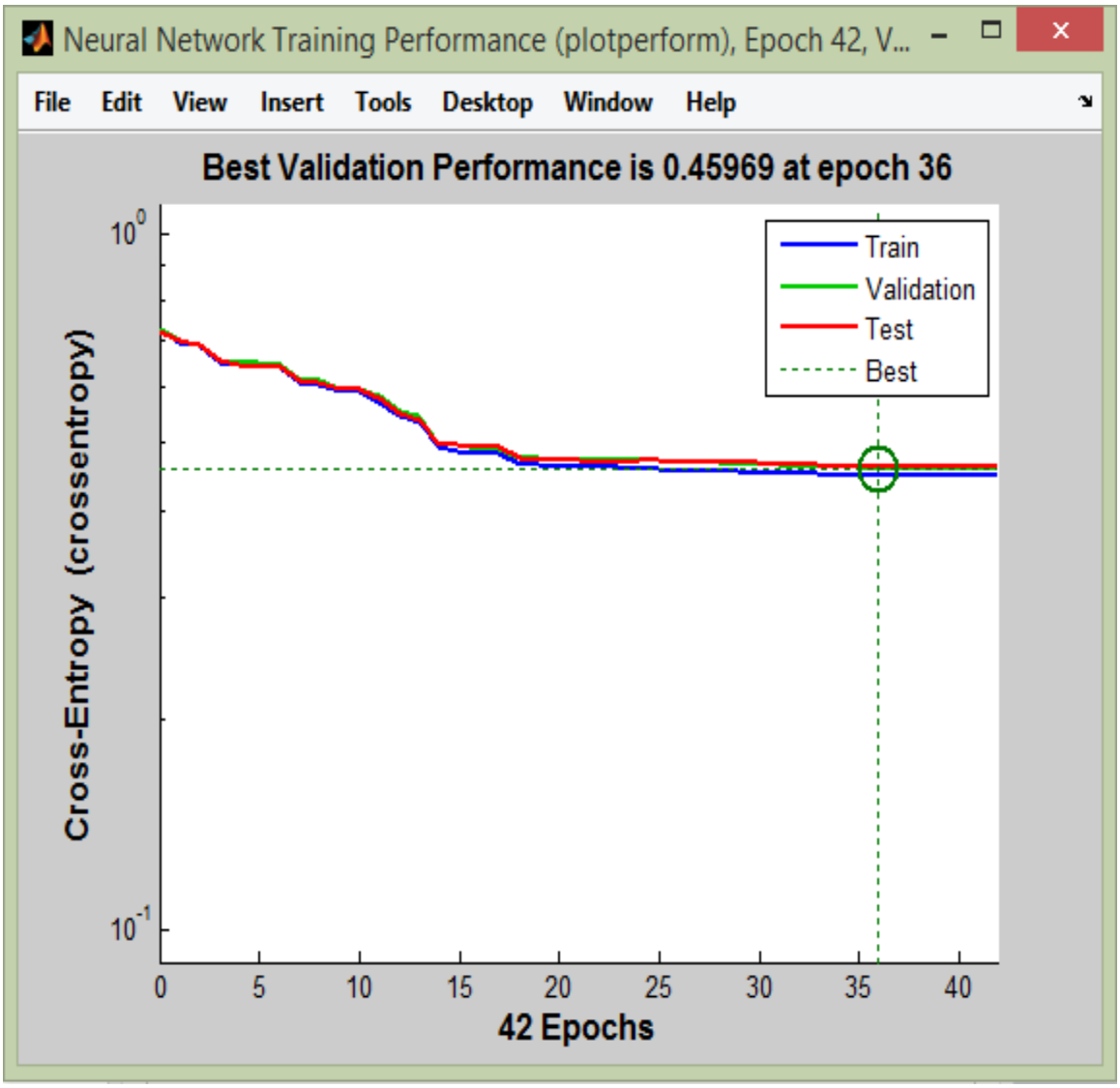
3.1 Αποτελέσματα εκπαίδευσης του τεχνητού νευρωνικού δικτύου στο περιβάλλον Matlab.

Στο εργαλείο Neural Net Pattern recognition του Matlab είχαμε τη δυνατότητα να επιλέξουμε τον αριθμό των νευρώνων του κρυφού επιπέδου. Τρέξαμε αρχικά το μοντέλο μας με αριθμό κρυφών νευρώνων ίσο με 4.Ορίσαμε να χρησιμοποιηθεί το 70% των δεδομένων ως δεδομένα εκπαίδευσης το 15% ως δεδομένα επικύρωσης και το υπόλοιπο 15% ως δεδομένα εκπαίδευσης. Το ΤΝΔ πέτυχε μετά από 42 επαναλήψεις ακρίβεια 80,4% στα δεδομένα εκπαίδευσης, 79.5% στα δεδομένα επικύρωσης και 79.8% στα δεδομένα δοκιμής. Τα ΤΝΔ εκπαιδεύτηκε με αλγόριθμο αντίστροφης τροφοδότησης τον Scaled conjugate gradient και το μέτρο απόδοσης του ήταν η διεντροπία (cross entropy). Παρακάτω φαίνονται ο πίνακας σύγχυσης με πληροφορίες για τις αληθώς και ψευδώς θετικές και αρνητικές ακολουθίες, η παράσταση Receiver Operating Characteristics curve (ROC – curve) και η γραφική παράσταση μείωσης της διεντροπίας.

Εικόνα 13. Πίνακας σύγχυσης για 4 νευρώνες.

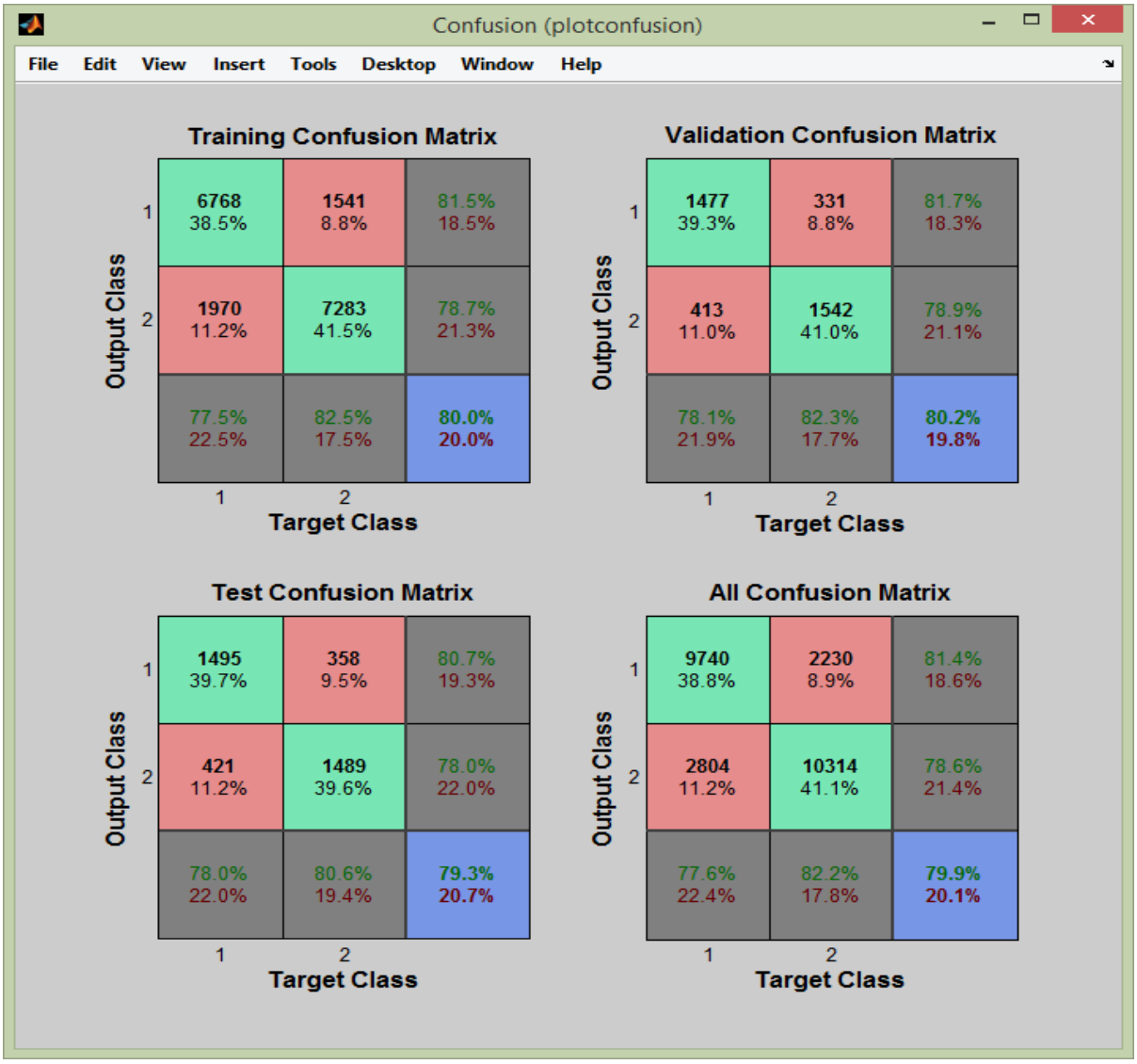


Εικόνα 14. Η παράσταση ROC (receiver operating characteristic) για 4 νευρώνες.

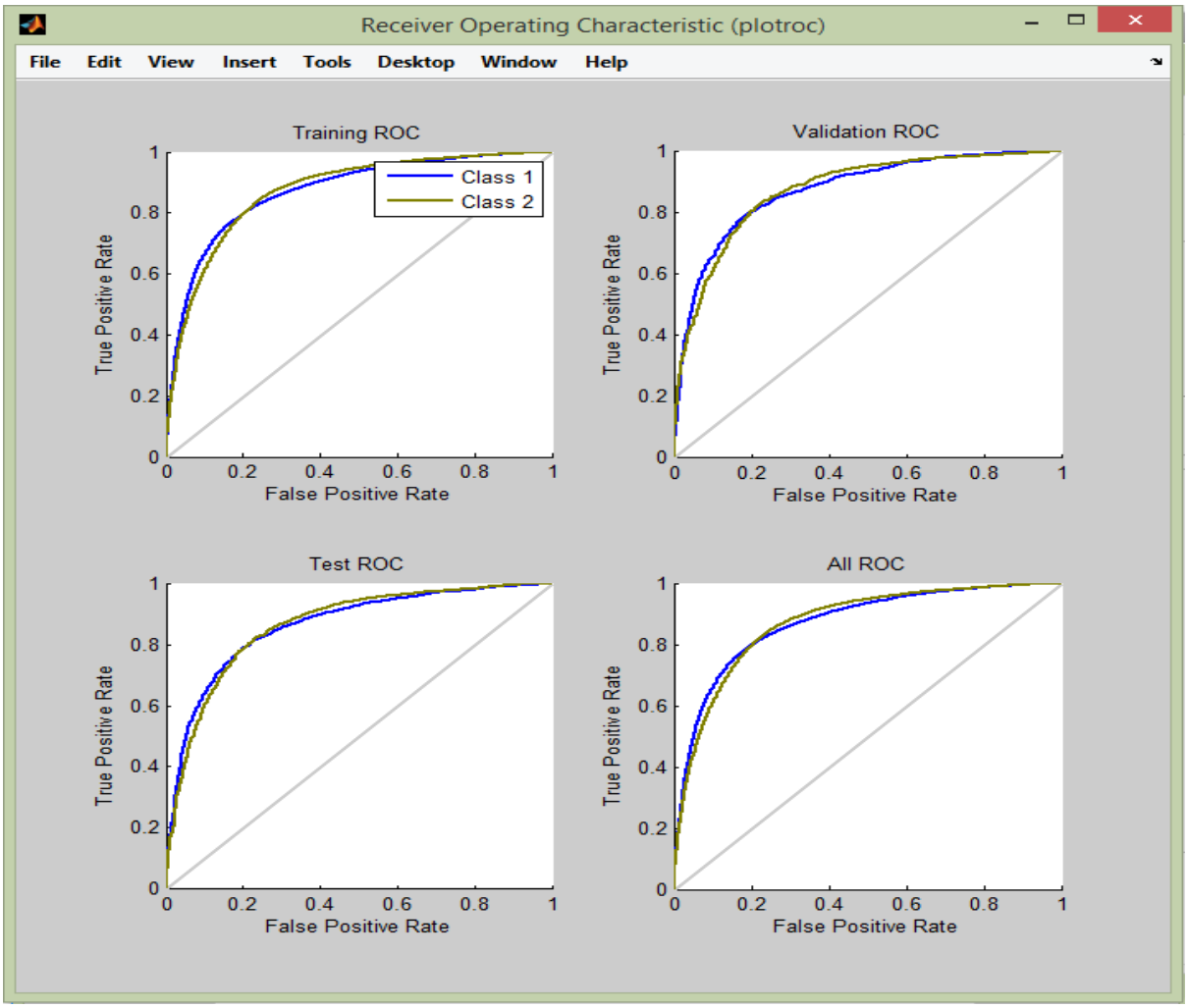


Εικόνα 15. Παράσταση μείωσης της διεντροπίας για 4 νευρώνες.

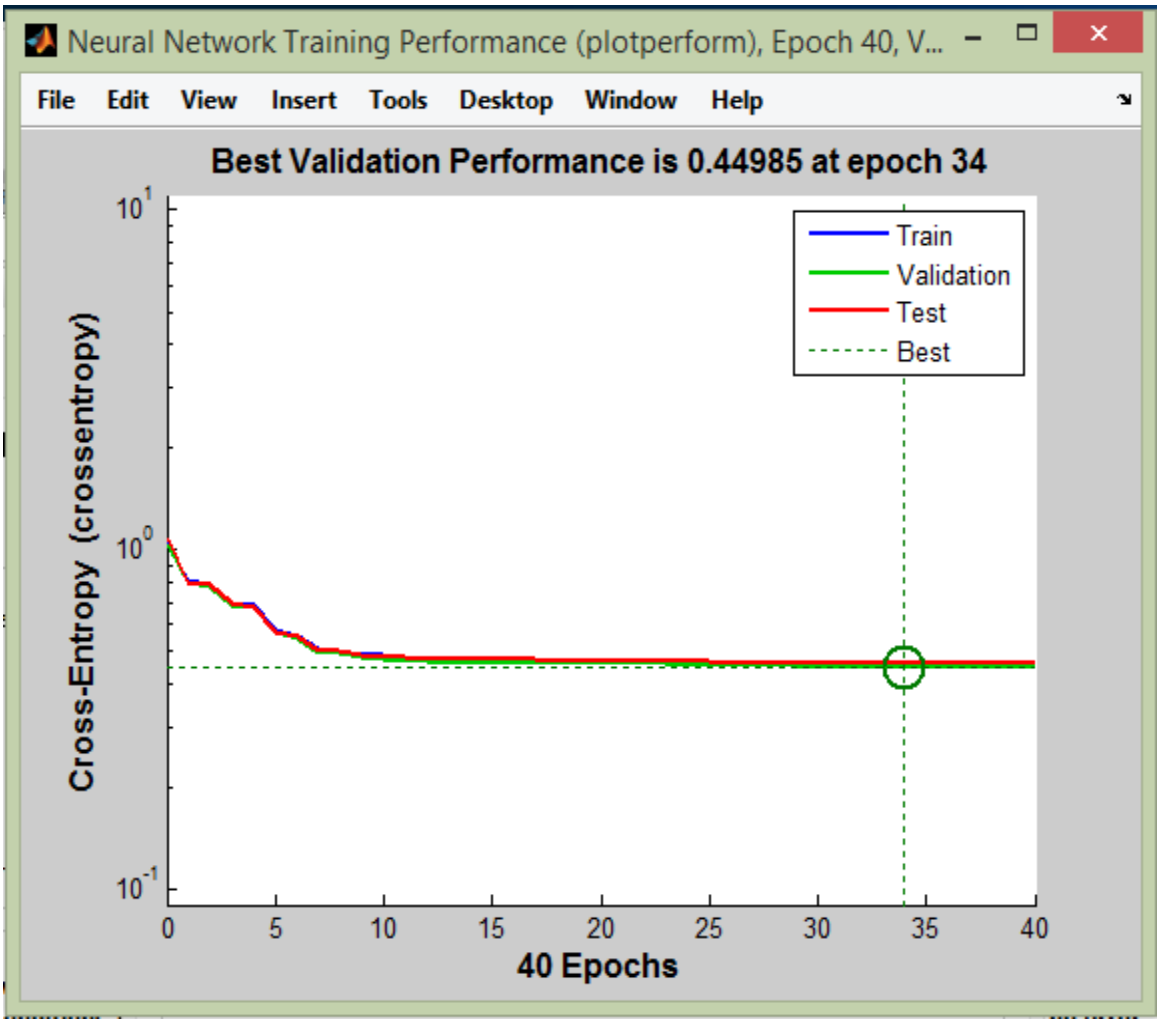
Τρέχοντας το δίκτυο με 11 κρυφούς νευρώνες αντί για 4 παρατηρήσαμε παρόμοια αποτελέσματα. Είχαμε μετά από 40 επαναλήψεις 80% ακρίβεια στα δεδομένα εκπαίδευσης, 80.2% στα δεδομένα επικύρωσης, και 79.3% στα δεδομένα δοκιμής. Παρακάτω παρατίθενται ο πίνακας σύγχυσης, η παράσταση ROC και η γραφική παράσταση μείωσης της διεντροπίας.



Εικόνα 16. Πίνακας σύγχυσης για 11 νευρώνες.



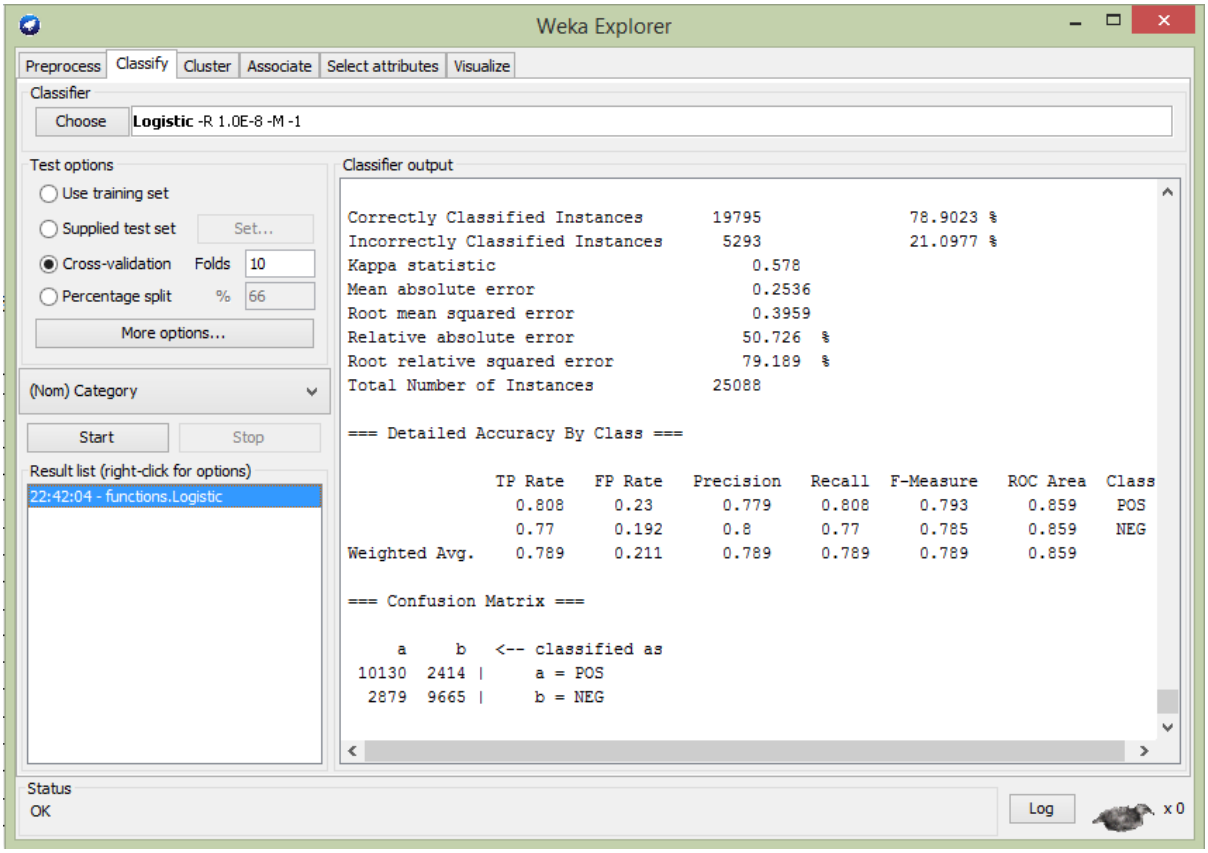
Εικόνα 17. Η παράσταση ROC για 11 νευρώνες.



Εικόνα 18. Η γραφική παράσταση μείωσης της διεντροπίας για 11 νευρώνες.

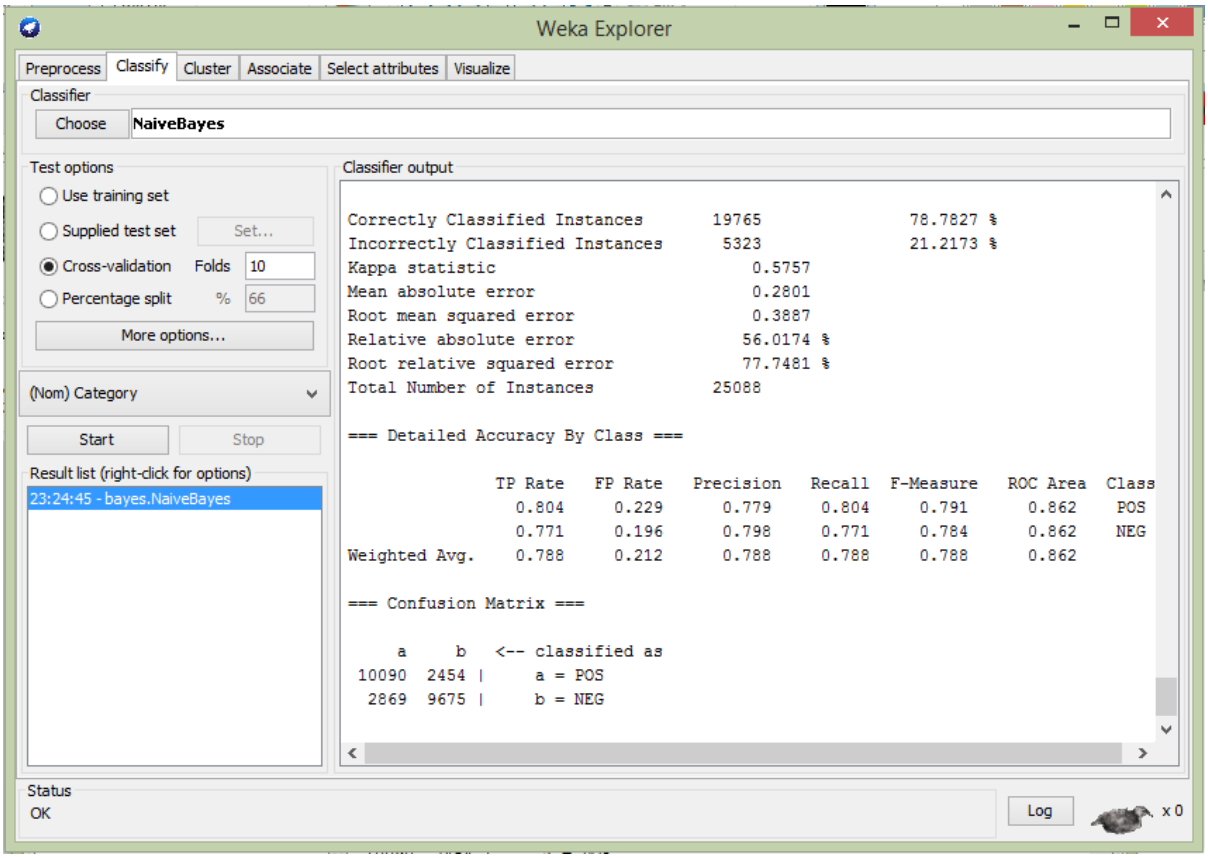
3.2 Αποτελέσματα ταξινόμησης του Weka

Αρχικά εισάγαμε τα δεδομένα μας με τη μορφή που προαναφέρθηκε στην ενότητα Υλικά και Μέθοδοι. Ο πρώτος αλγόριθμος που τρέξαμε στο Weka ήταν αυτός της λογιστικής παλινδρόμησης. Επιλέξαμε τον αλγόριθμο Logistic από το παράθυρο επιλογής αλγορίθμων. Ο αλγόριθμος πέτυχε 78.9023% ακρίβεια..



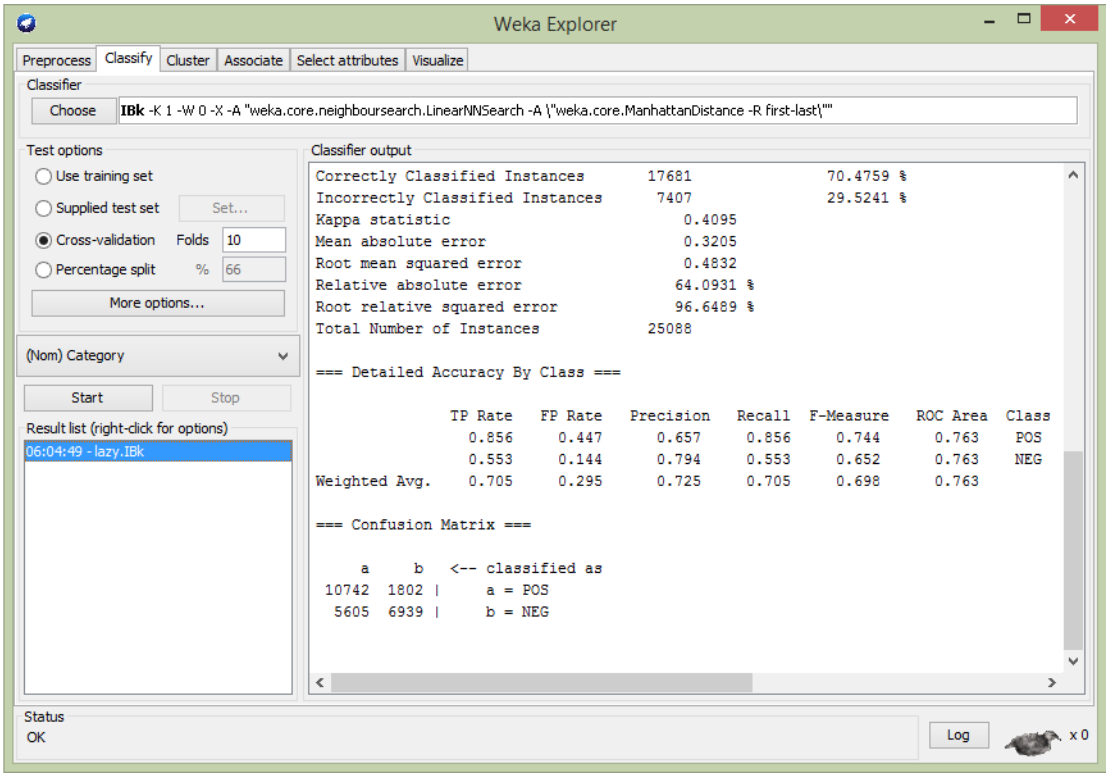
Εικόνα 19. Αποτελέσματα λογιστικής παλινδρόμησης στο WEKA.

Ο επόμενος αλγόριθμος που επιλέξαμε ήταν ο NaiveBayes. Η ακρίβεια αυτού του αλγορίθμου ήταν 78.7827%. Το True Positive rate για τις θετικές ακολουθίες ήταν 0.804 ενώ για τις αρνητικές 0.771.



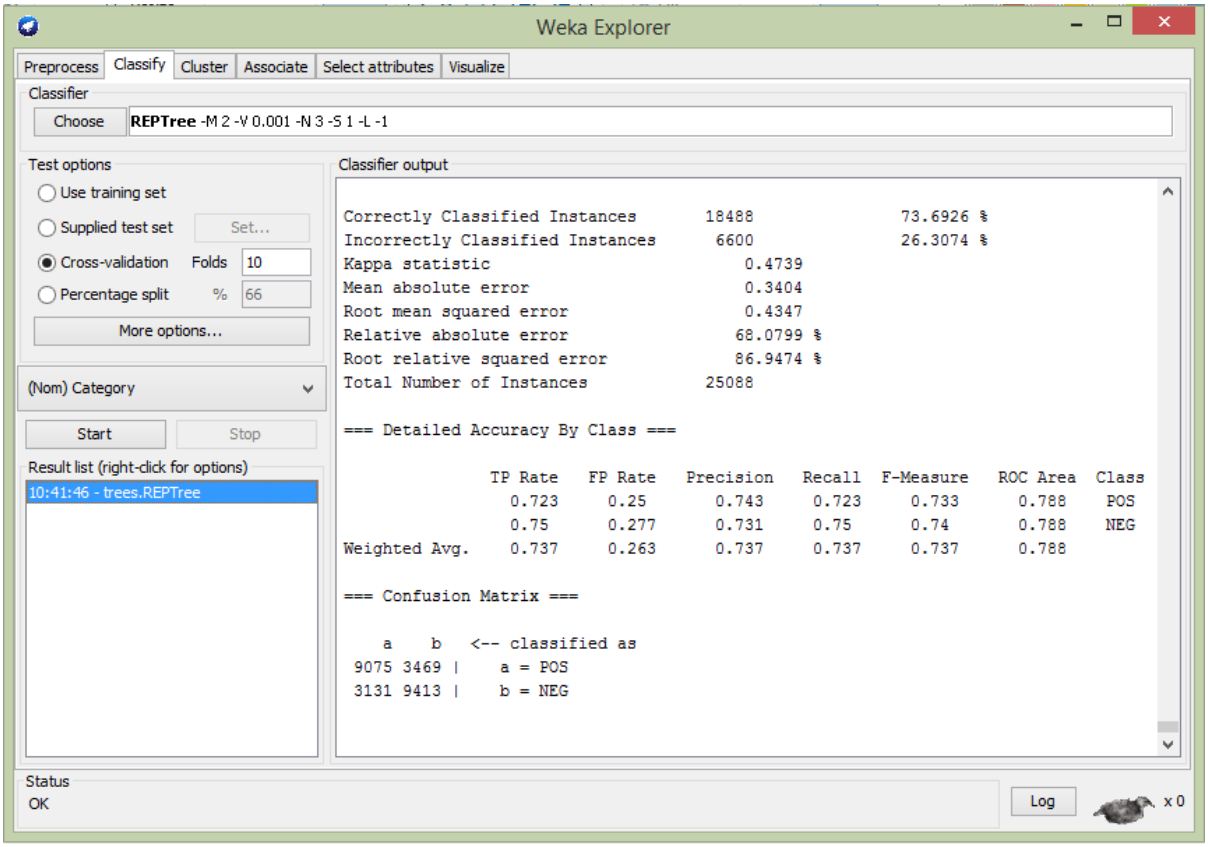
Εικόνα 20.Αποτελέσματα ταξινόμησης του Naive Bayes στο WEKA.

Έπειτα τρέξαμε τον αλγόριθμο K-nearest neighbours. Ο αλγόριθμος αυτός χρησιμοποιεί ως κριτήριο ταξινόμησης την ομοιότητα μεταξύ των ακολουθιών. Ρυθμίζοντας την τιμή κ στο παράθυρο των ρυθμίσεων ορίζουμε με βάση πόσες ομοιότερες ακολουθίες να γίνει η πρόβλεψη. Έχουμε επίσης τη δυνατότητα να ορίσουμε το είδος της απόστασης που θα χρησιμοποιεί ο αλγόριθμος. Τρέξαμε τον αλγόριθμο με cross validation: True και απόσταση Manhattan και είχαμε ακρίβεια 70.4759% με True Positive rate 0.856 για τις θετικές και 0.553 για τις αρνητικές ακολουθίες.



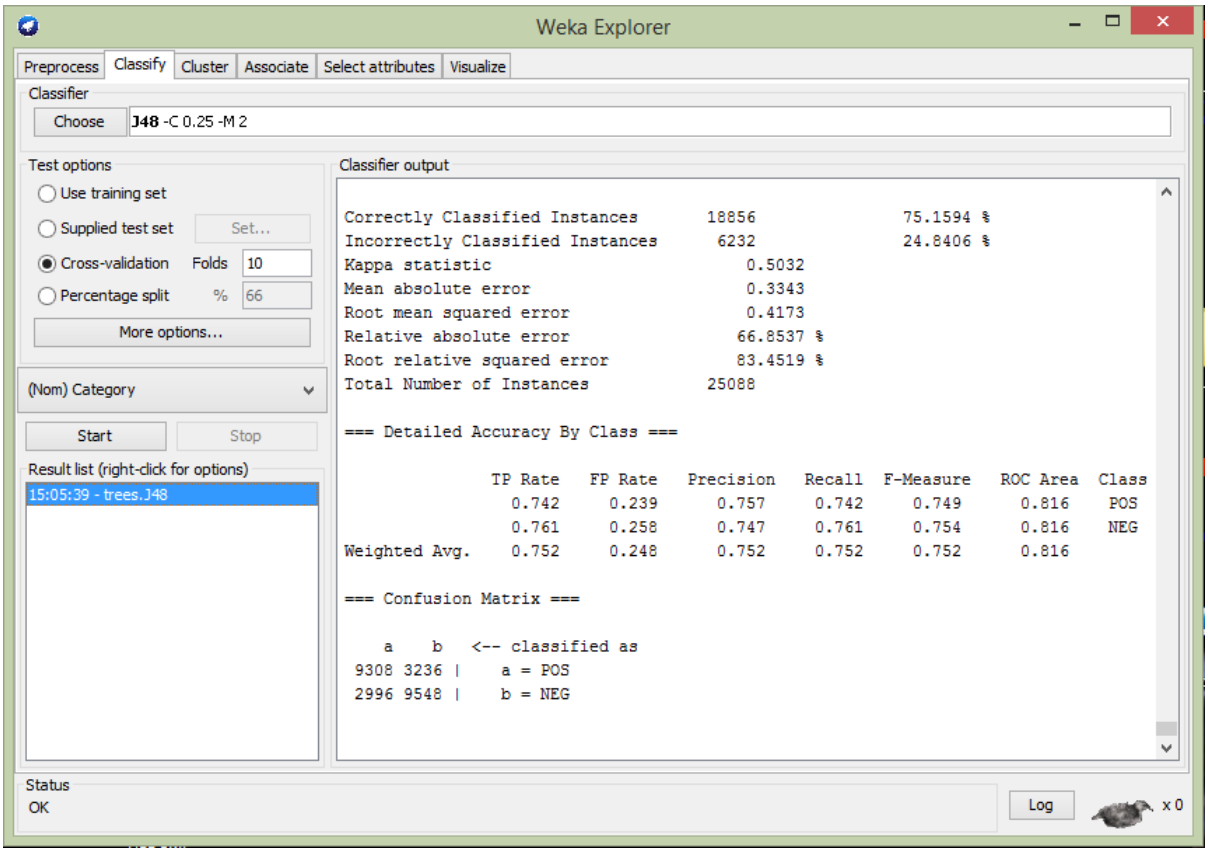
Εικόνα 21. Αποτελέσματα Ανάλυσης του ταξινομητή K-nearest neighbours στο WEKA.

Ο επόμενος αλγόριθμος που τρέξαμε ήταν ο αλγόριθμος δέντρων αποφάσεων REPTree. Ο αλγόριθμος είχε ακρίβεια 73.6926 % με True Positive rate 0.723 για τις θετικές και 0.75 για τις αρνητικές ακολουθίες.



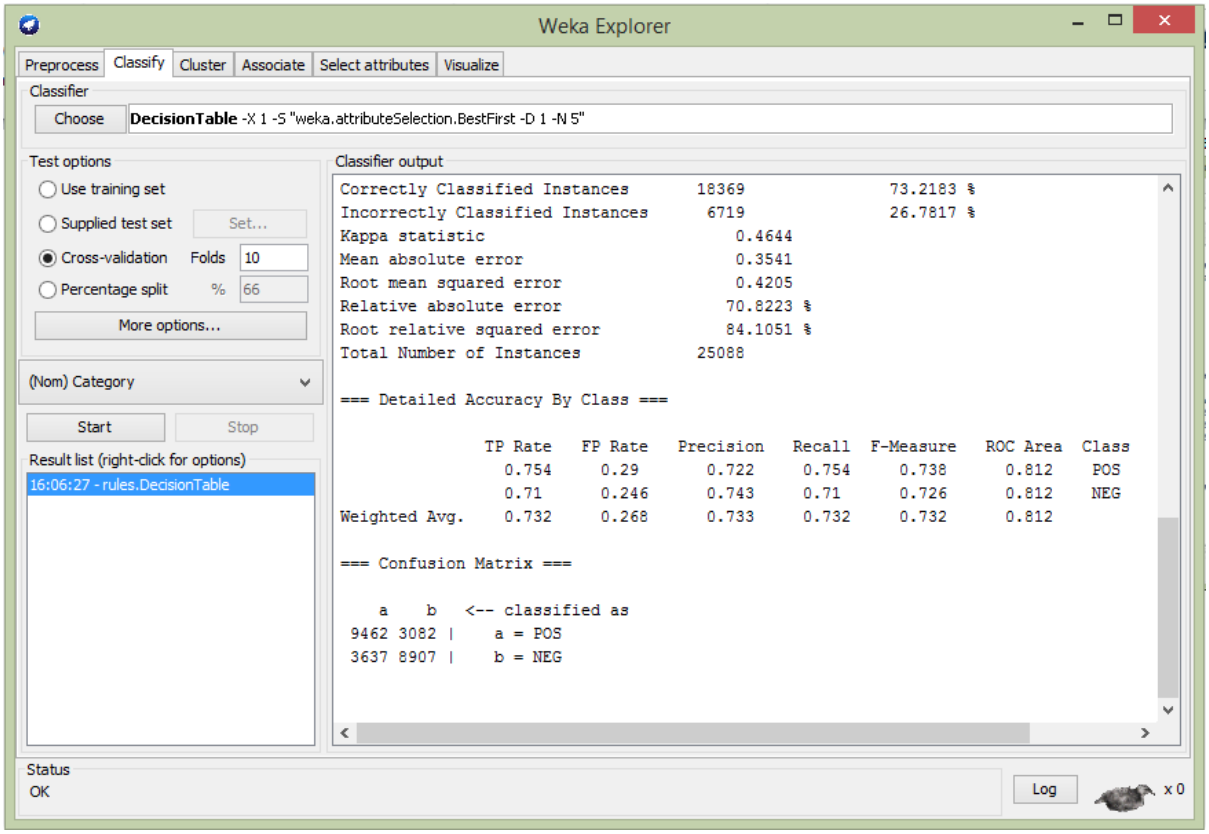
Εικόνα 22. Αποτελέσματα ταξινόμησης του αλγορίθμου Reptree στο WEKA.

Ο επόμενος αλγόριθμος που τρέξαμε ήταν ο αλγόριθμος δέντρων αποφάσεων J48.Ο αλγόριθμος είχε ακρίβεια 75.1594% με True Positive rate 0.742 για τις θετικές και 0.761 για τις αρνητικές ακολουθίες.



Εικόνα 23. Αποτελέσματα ταξινόμησης του αλγορίθμου J48 στο WEKA.

Έπειτα τρέξαμε τον αλγόριθμο Decision Table. Ο αλγόριθμος είχε ακρίβεια 73.2183% με True Positive rate 0.754 για τις θετικές και 0.71 για τις αρνητικές ακολουθίες.



Εικόνα 24. Αποτελέσματα ταξινόμησης του αλγορίθμου DecisionTable στο WEKA.

Τα αποτελέσματα συνοψίζονται στον παρακάτω πίνακα.

Πίνακας 3. Η ακρίβεια πρόβλεψης των θέσεων φωσφορυλίωσης στον αρουραίο για 7 αλγορίθμους ταξινόμησης του WEKA.

Αλγόριθμος Ταξινόμησης	Ακρίβεια	True Positive Rate(POS)	True Positive Rate(NEG)
Λογιστική Παλινδρόμηση	78.9023%	0.808	0.77
Naive Bayes	78.7827%	0.804	0.771
K-nearest neighbours	70.4759%	0.856	0.553
REPTree	73.6926 %	0.723	0.75
J48	75.1594 %	0.742	0.761
Decision Table	73.2183 %	0.754	0.71

3.3 Αποτελέσματα εκπαίδευσης του τεχνητού νευρωνικού δικτύου στο Keras/Tensorflow.

Αρχικά, δημιουργήσαμε ένα μοντέλο με 2 κρυφά επίπεδα των 8 νευρώνων το καθένα και 1 επίπεδο εξόδου ενεργοποιούμενο με σιγμοειδή συνάρτηση. Σαν συνάρτηση απώλειας ορίσαμε τη δυαδική διεντροπία (binary crossentropy), σαν βελτιστοποιητή ορίσαμε τον αλγόριθμο adam και ως metrics θέσαμε την ακρίβεια. Ως συνάρτηση ενεργοποίησης θέσαμε τη συνάρτηση μονάδων γραμμικής ανόρθωσης (relu). Τρέξαμε το μοντέλο για 200 εποχές και χρησιμοποιήσαμε το 80% των δεδομένων για εκπαίδευση και επικύρωση ενώ το υπόλοιπο 20% για δοκιμή.

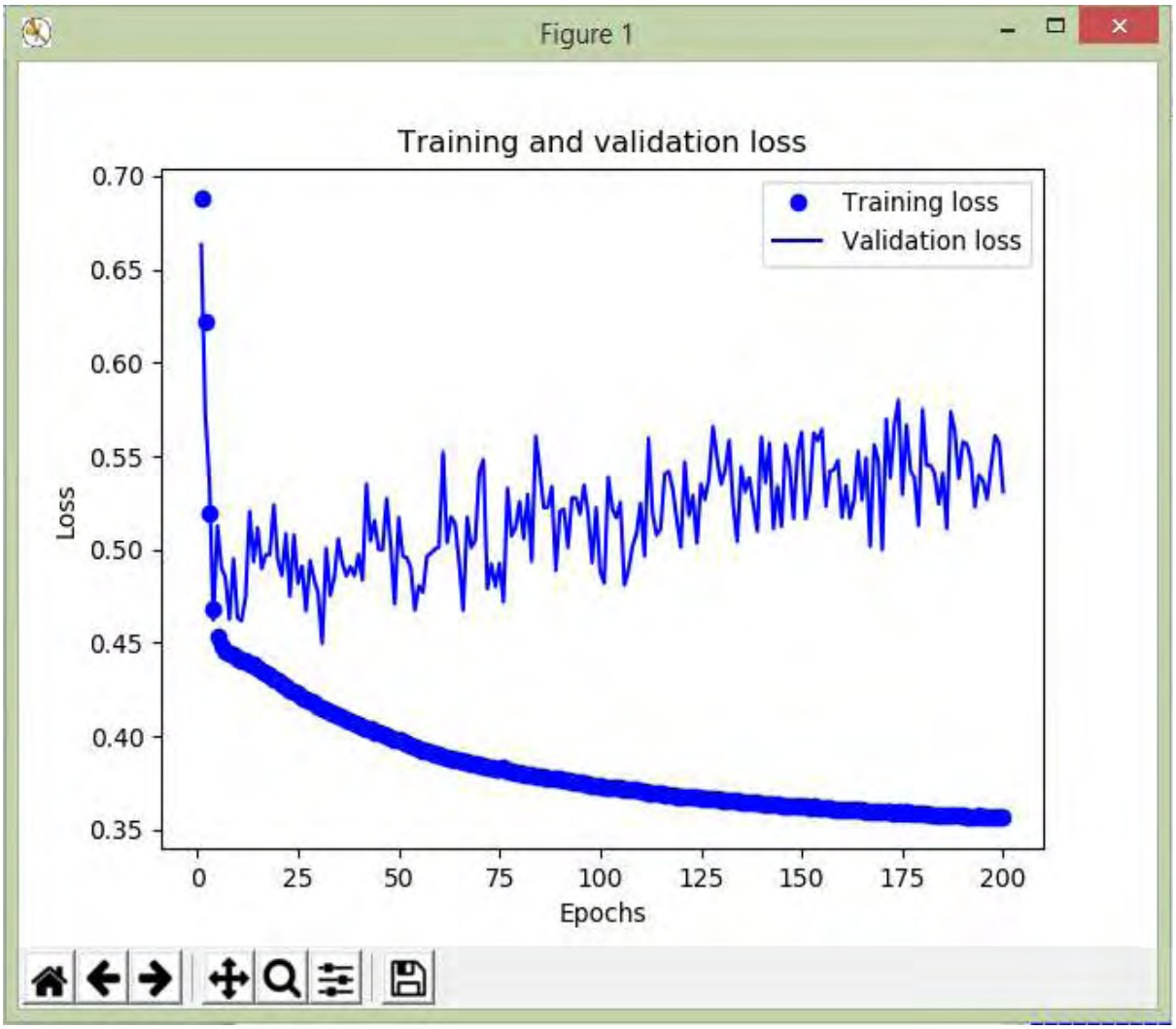
Το μοντέλο πέτυχε ακρίβεια 77.4%. Παρατίθενται η σύνοψη του μοντέλου, τα αποτελέσματα με την ακρίβεια και τον πίνακα σύγχυσης και οι γραφικές παραστάσεις της απώλειας και της ακριβείας στα δεδομένα εκπαίδευσης και επικύρωσης.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 8)                 1768
_________________________________________________________________
dense_2 (Dense)              (None, 8)                 72
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 9
=================================================================
Total params: 1,849
Trainable params: 1,849
Non-trainable params: 0
```

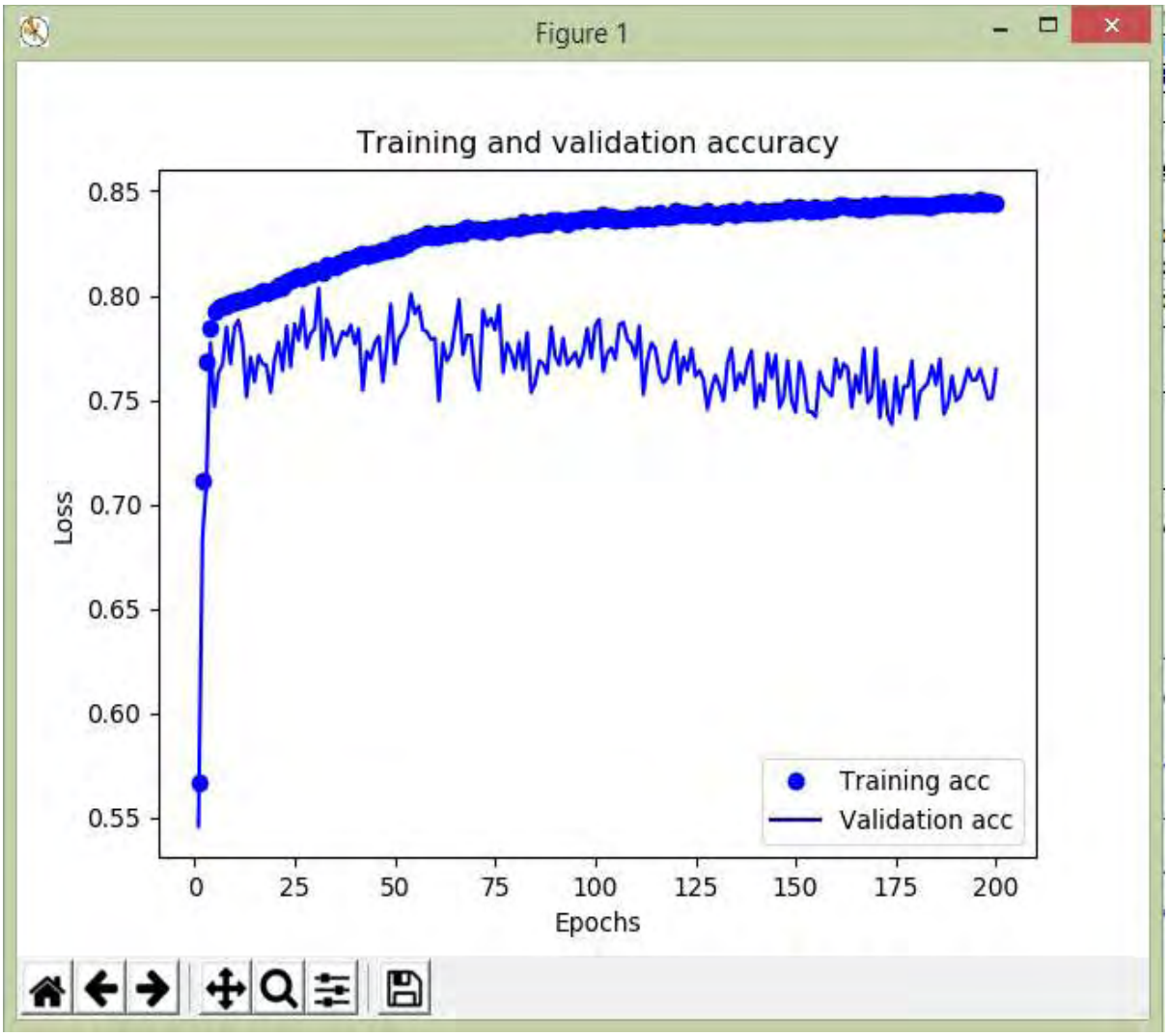
Εικόνα 25. Η σύνοψη του μοντέλου δύο κρυφών επιπέδων χωρίς Dropout.

```
[[1820  688]
 [ 444 2064]]
Accuracy: 0.7743221690590112
0.7743221690590112
```

Εικόνα 26. Ο πίνακας σύγχυσης και η ακρίβεια του μοντέλου δύο κρυφών επιπέδων χωρίς Dropout.



Εικόνα 27. Οι γραφικές παραστάσεις της απώλειας του μοντέλου δύο κρυφών επιπέδων χωρίς Dropout.



Εικόνα 28. Οι γραφικές παραστάσεις της ακρίβειας του μοντέλου δύο κρυφών επιπέδων χωρίς Dropout.

Παρατηρούμε από τις παραπάνω γραφικές παραστάσεις ότι μετά από κάποιες εποχές, ενώ η απώλεια μειώνεται και η ακρίβεια αυξάνεται στα δεδομένα εκπαίδευσης, το αντίθετο συμβαίνει για τα δεδομένα επικύρωσης. Το μοντέλο ενώ πετυχαίνει καλή ακρίβεια πρόβλεψης στα δεδομένα εκπαίδευσης, έχει χάσει την ικανότητα γενίκευσης για τα δεδομένα επικύρωσης. Παρουσιάζει δηλαδή το λεγόμενο overfit.

Τρέξαμε ξανά το ίδιο μοντέλο αυτή τη φορά χρησιμοποιώντας στρατηγικές καταπολέμησης του overfit. Χρησιμοποιήσαμε την τεχνική Dropout με τιμή του Dropout ίση με 0.2 και την τεχνική της κανονικοποίησης των βαρών (weight regularization). Η τεχνική Dropout απορρίπτει τυχαία κάποια χαρακτηριστικά εξόδου των κρυφών επιπέδων αποτρέποντας το ΤΝΔ να απομνημονεύσει μη σημαντικά μοτίβα. Η τεχνική της κανονικοποίησης των βαρών περιορίζει την πολυπλοκότητα του μοντέλου αναγκάζοντας τα βάρη να πάρουν μικρότερες τιμές και κάνοντας την κατανομή τους περισσότερο κανονική.

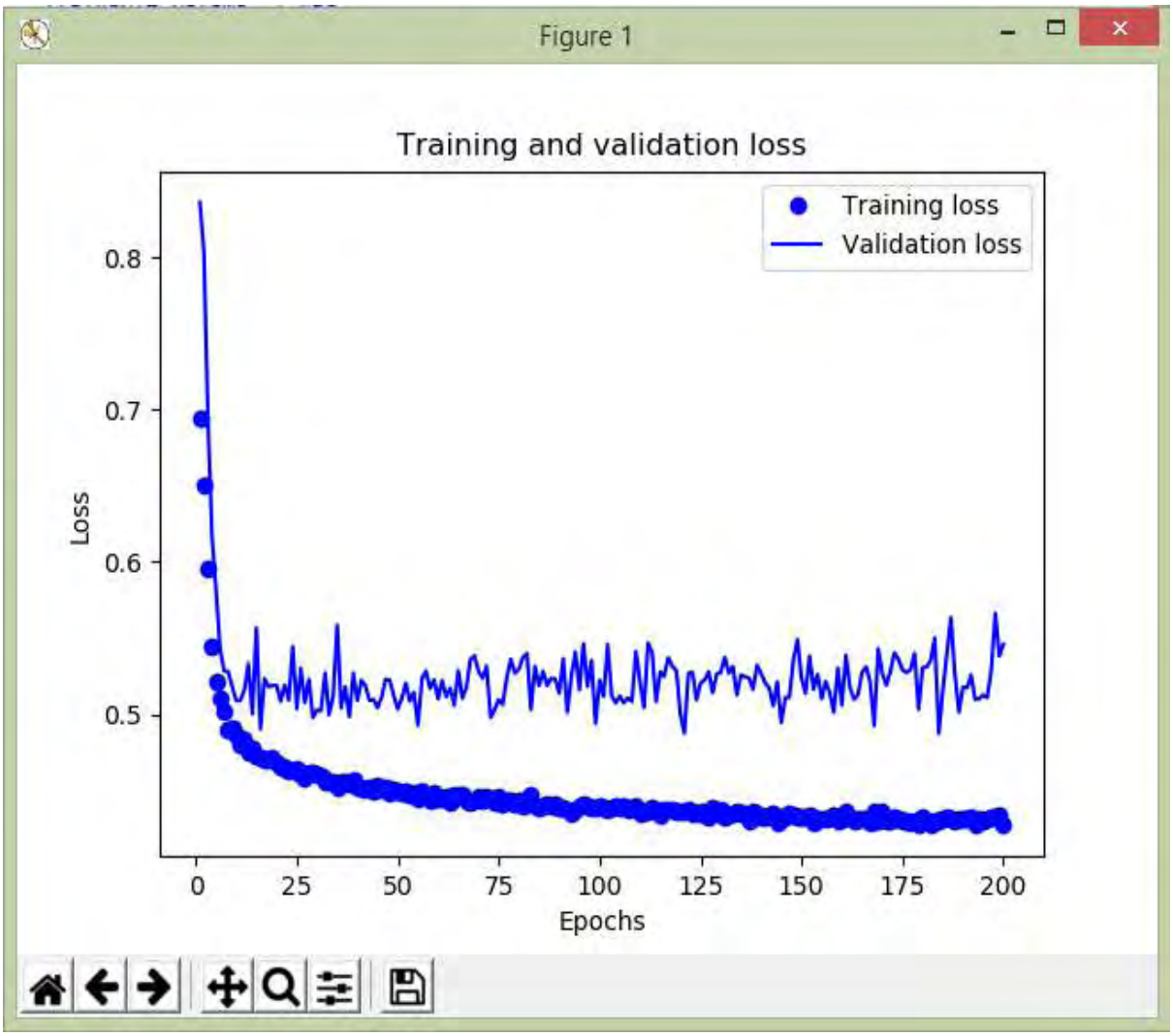
Το μοντέλο μετά τη χρήση των παραπάνω στρατηγικών είχε ακρίβεια 78.1%. Παρακάτω βλέπουμε τα αποτελέσματα με την ακρίβεια και τον πίνακα σύγχυσης και τις γραφικές παραστάσεις της απώλειας και της ακριβείας στα δεδομένα εκπαίδευσης και επικύρωσης.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 8)                 1768
_________________________________________________________________
dropout_1 (Dropout)          (None, 8)                 0
_________________________________________________________________
dense_2 (Dense)              (None, 8)                 72
_________________________________________________________________
dropout_2 (Dropout)          (None, 8)                 0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 9
=================================================================
Total params: 1,849
Trainable params: 1,849
Non-trainable params: 0
```

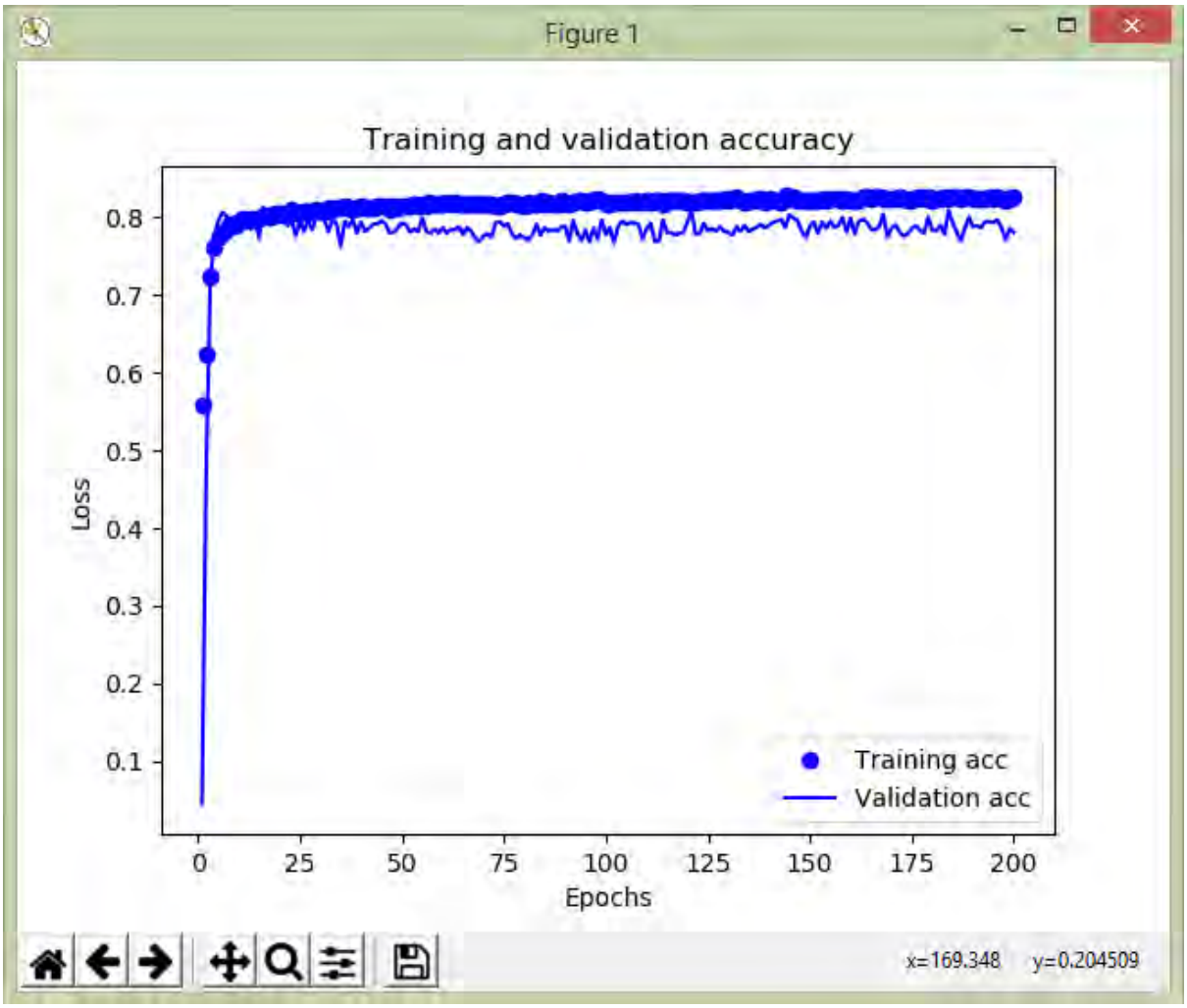
Εικόνα 29. Η σύνοψη του μοντέλου του μοντέλου δύο κρυφών επιπέδων με Dropout 0.2 .

```
[[1823  685]
 [ 411 2097]]
Accuracy: 0.7814992025518341
0.7814992025518341
```

Εικόνα 30. Ο πίνακας σύγχυσης και η ακρίβεια του μοντέλου δύο κρυφών επιπέδων με Dropout 0.2.



Εικόνα 31. Οι γραφικές παραστάσεις της απώλειας του μοντέλου δύο κρυφών επιπέδων με Dropout 0.2.

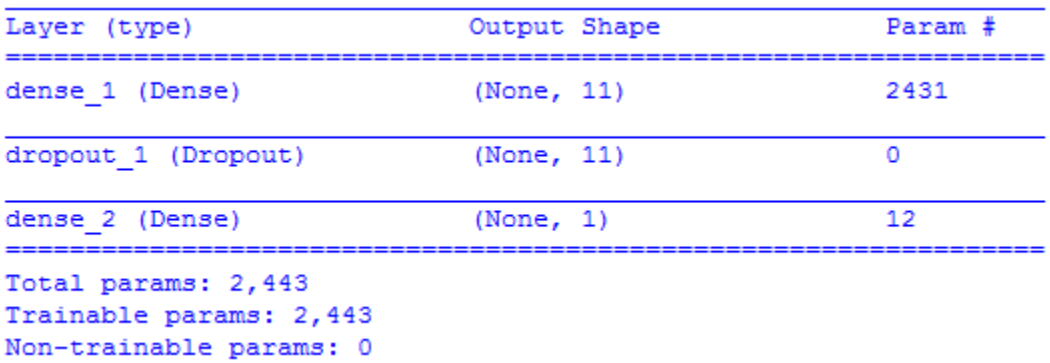


Εικόνα 32. Οι γραφικές παραστάσεις της ακρίβειας του μοντέλου δύο κρυφών επιπέδων με Dropout 0.2.

Συγκρίνοντας τα δύο πειράματα, παρατηρούμε ότι το φαινόμενο του overfit έχει μειωθεί με την χρήση του Dropout. Η υψηλότερη ακρίβεια στα δεδομένα εκπαίδευσης που καταγράφεται στο πρώτο πείραμα σημαίνει ότι το μοντέλο απομνημονεύει τα δεδομένα εκπαίδευσης χάνοντας όμως την ικανότητα γενίκευσης.

Έπειτα, δημιουργήσαμε ένα μοντέλο ενός κρυφού επιπέδου και δοκιμάσαμε διαφορετικούς συνδυασμούς αριθμού νευρώνων και τιμών Dropout. Δημιουργήσαμε μοντέλα των 8, 11 και 14 νευρώνων θέτοντας εναλλάξ για το καθένα τιμές Dropout ίσες με 0, 0.2 και 0.4.

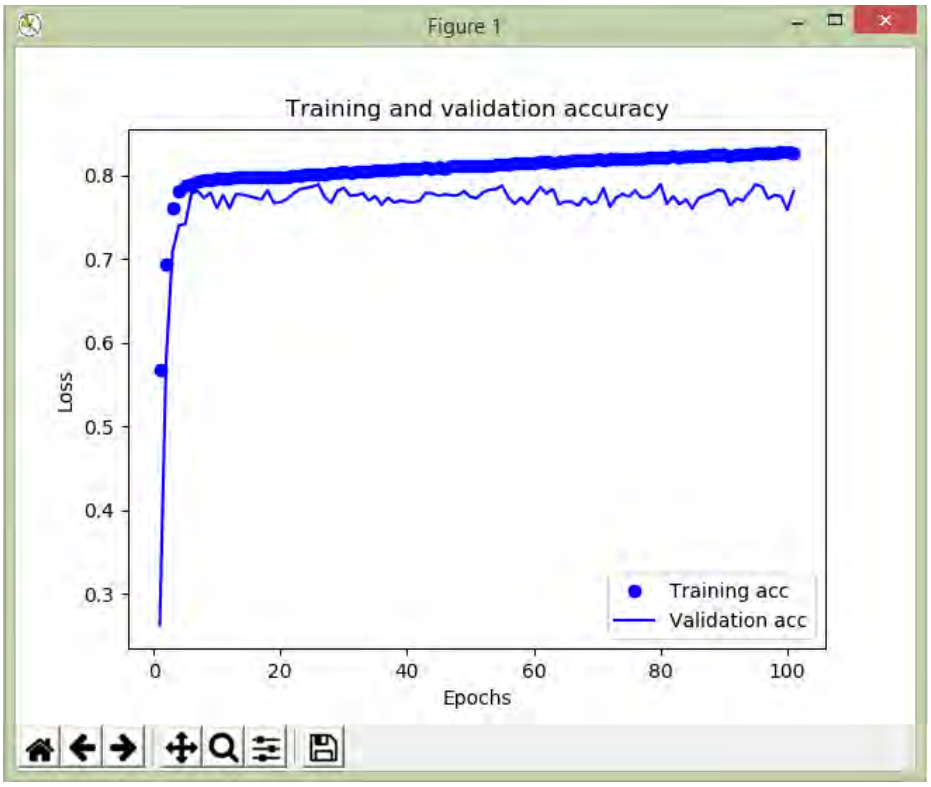
Παρακάτω, παρατίθεται η σύνοψη της αρχιτεκτονικής του ΤΝΔ για 11 κρυφούς νευρώνες, ο πίνακας των αποτελεσμάτων για κάθε περίπτωση (Πίνακας 4), καθώς και οι γραφικές παραστάσεις της ακρίβειας.



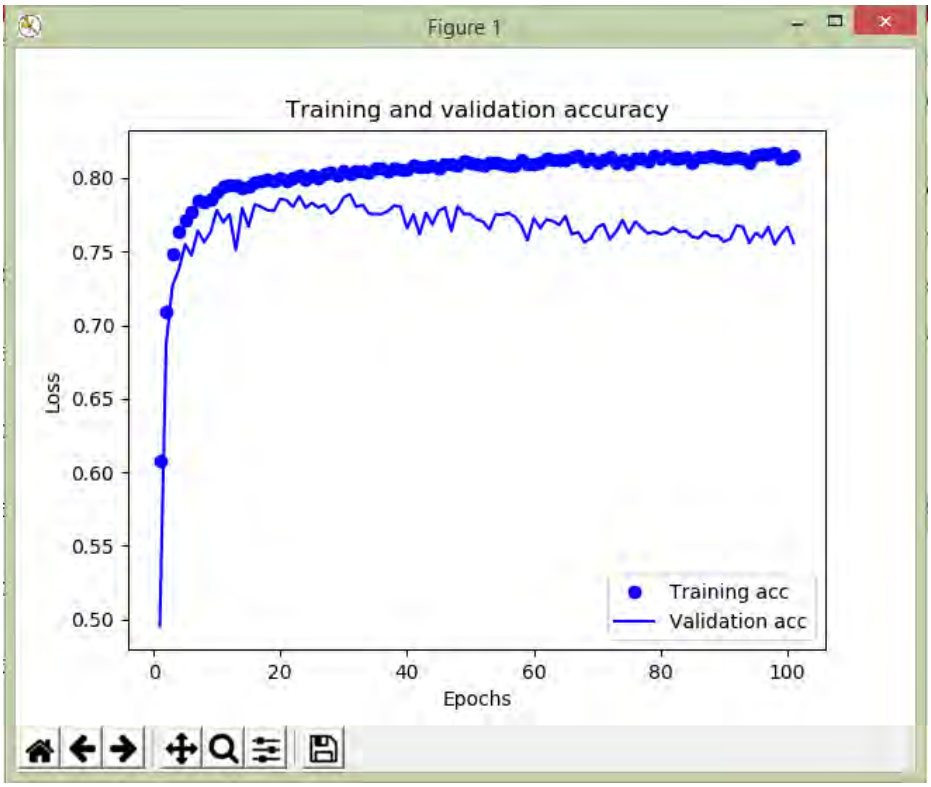
Εικόνα 33. Η σύνοψη του μοντέλου για 11 κρυφούς νευρώνες.

Πίνακας 4. Η ακρίβεια για κάθε συνδυασμό κρυφών νευρώνων και Dropout.

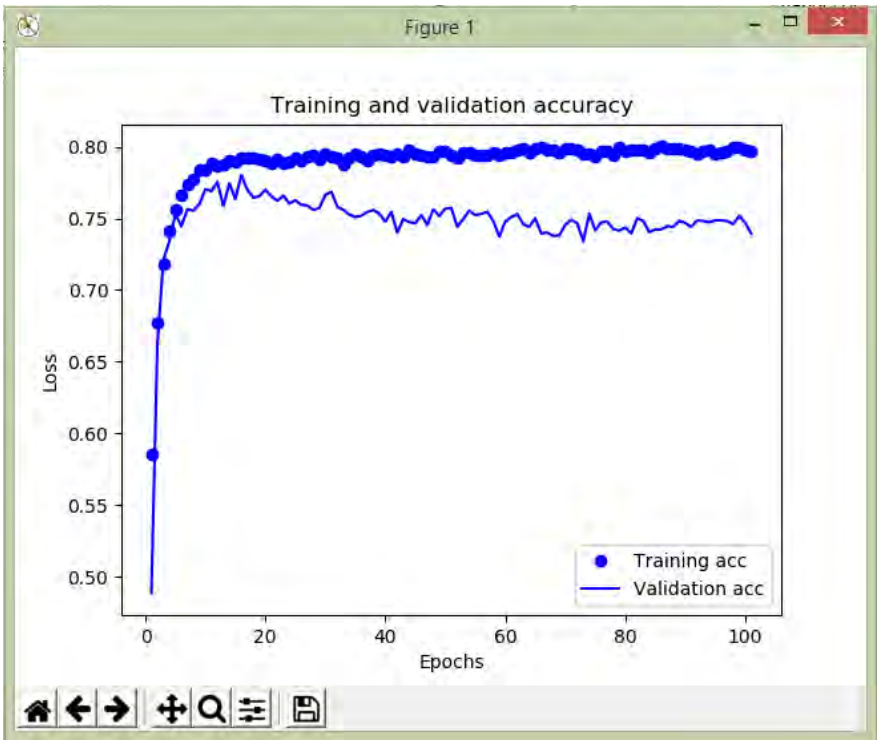
Κρ.Νευρ./ Dropout	0	0.2	0.4	
8	78.50%	77.60%	77.60%	
11	77.80%	78.20%	78.10%	
14	77.50%	78.60%	78.60%	



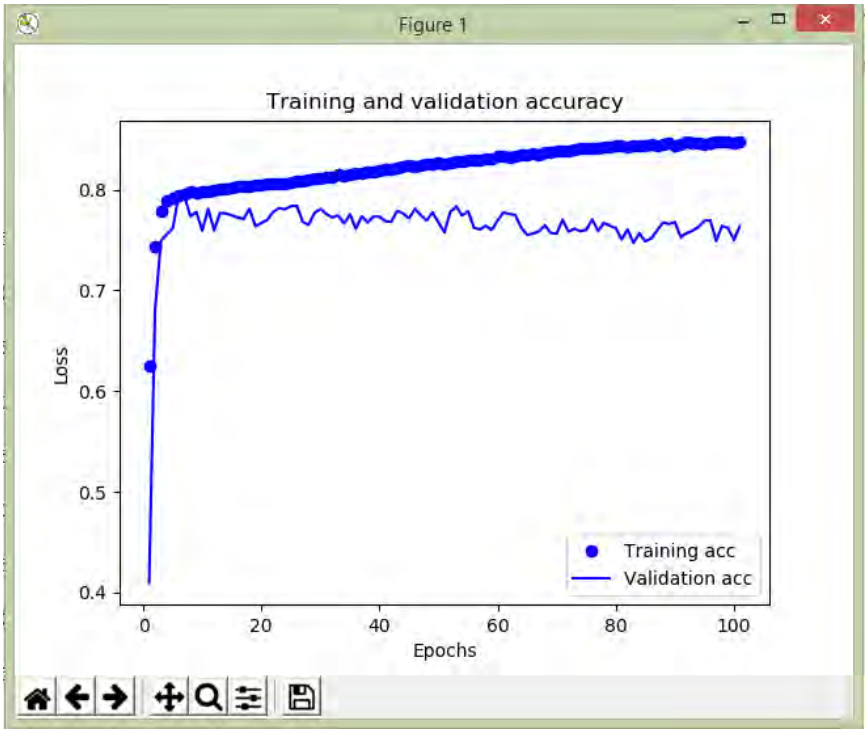
Εικόνα 34. Οι γραφικές παραστάσεις της ακρίβειας για 8 κρυφούς νευρώνες και Dropout ίσο με 0.



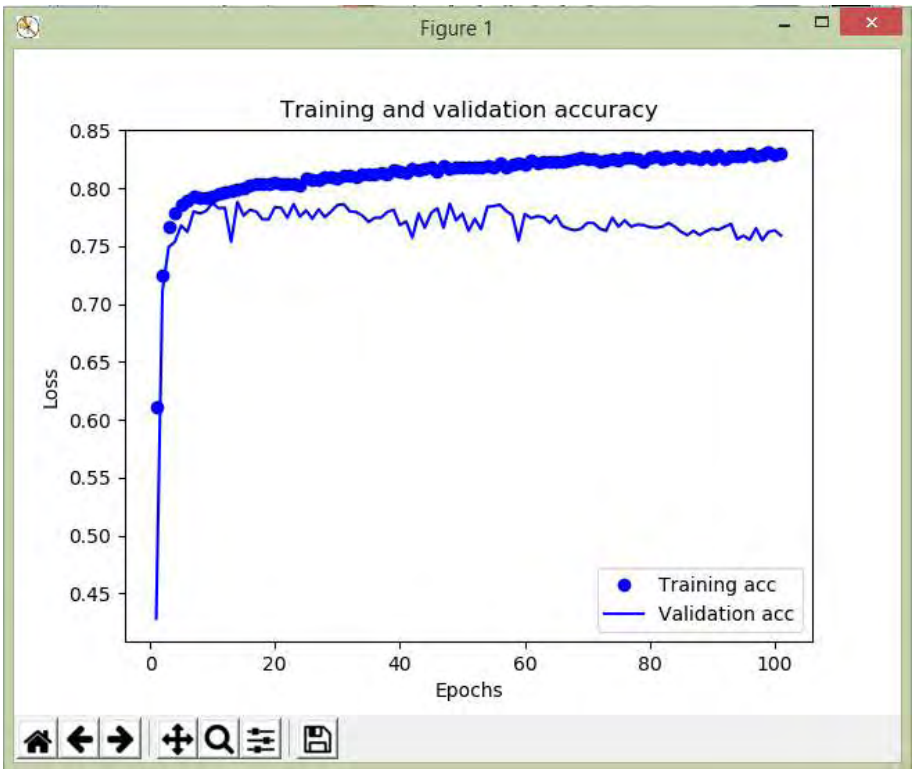
Εικόνα 35. Οι γραφικές παραστάσεις της ακρίβειας για 8 κρυφούς νευρώνες και Dropout ίσο με 0.2.



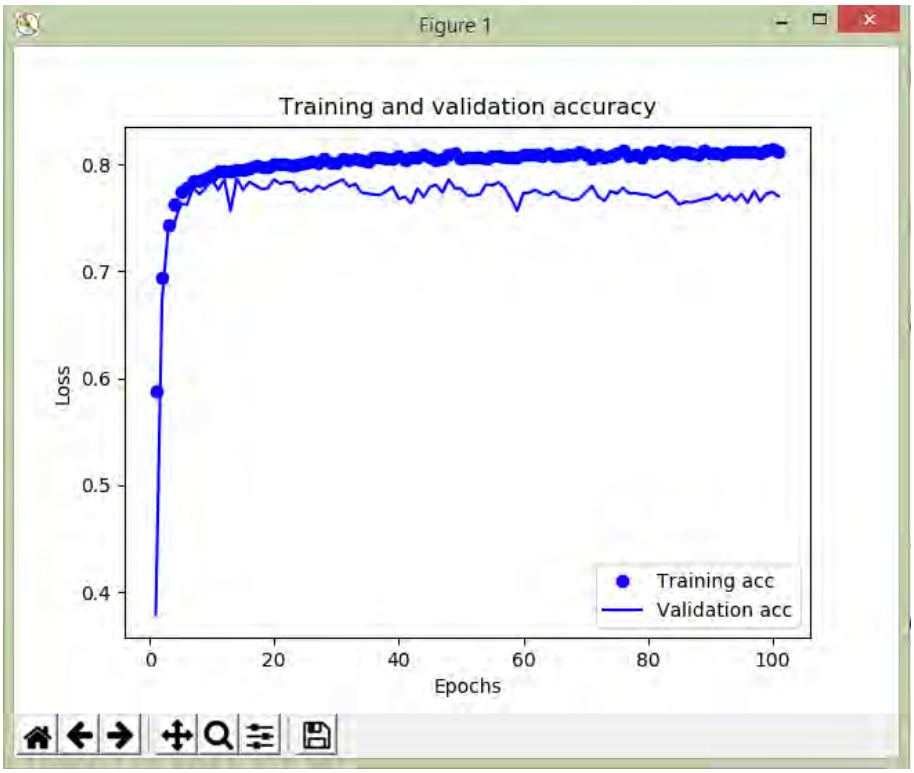
Εικόνα 36. Οι γραφικές παραστάσεις της ακρίβειας για 8 κρυφούς νευρώνες και Dropout ίσο με 0.4.



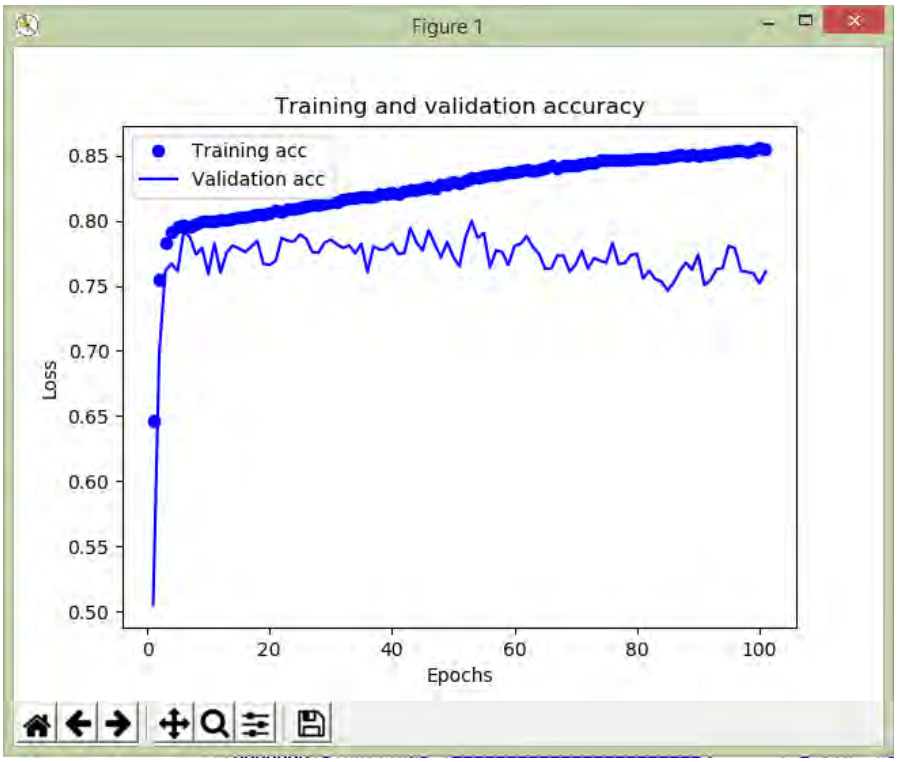
Εικόνα 37. Οι γραφικές παραστάσεις της ακρίβειας για 11 κρυφούς νευρώνες και Dropout ίσο με 0.



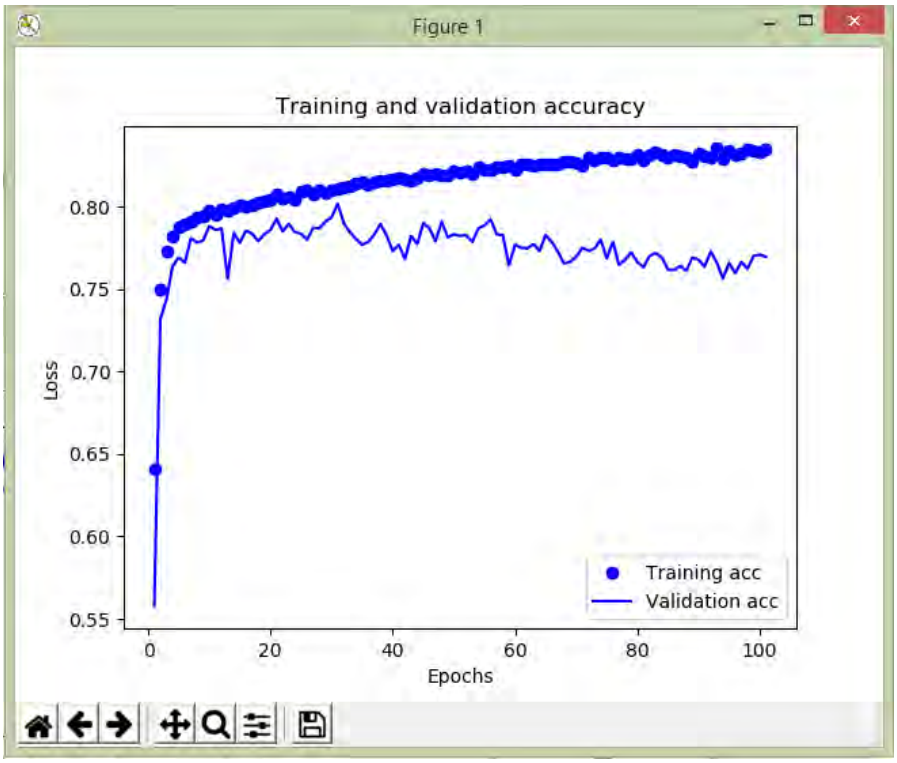
Εικόνα 38. Οι γραφικές παραστάσεις της ακρίβειας για 11 κρυφούς νευρώνες και Dropout ίσο με 0.2.



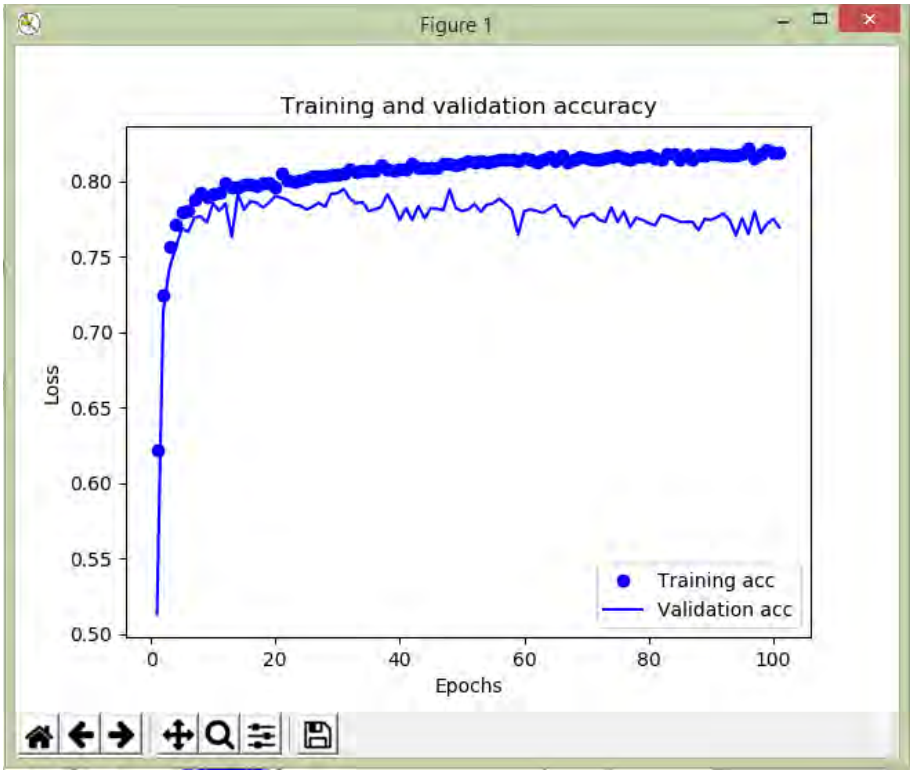
Εικόνα 39. Οι γραφικές παραστάσεις της ακρίβειας για 11 κρυφούς νευρώνες και Dropout ίσο με 0.4.



Εικόνα 40. Οι γραφικές παραστάσεις της ακρίβειας για 14 κρυφούς νευρώνες και Dropout ίσο με 0.



Εικόνα 41. Οι γραφικές παραστάσεις της ακρίβειας για 14 κρυφούς νευρώνες και Dropout ίσο με 0.2.



Εικόνα 42. Οι γραφικές παραστάσεις της ακρίβειας για 14 κρυφούς νευρώνες και Dropout ίσο με 0.4.

Παρατηρώντας τα παραπάνω αποτελέσματα βλέπουμε ότι το κέρδος στην ακρίβεια από την αύξηση του αριθμού των κρυφών νευρών είναι μικρό. Η εφαρμογή του Dropout γίνεται σημαντική όσο αυξάνεται ο αριθμός των νευρώνων, όσο αυξάνονται δηλαδή οι παράμετροι του μοντέλου. Για μικρό αριθμό νευρώνων η επίδραση του Dropout δεν είναι τόσο εμφανής, για μεγαλύτερο όμως αριθμό νευρώνων η χρήση του Dropout είναι απαραίτητη.

Συμπεράσματα

Στην παρούσα διπλωματική εργασία, εκπαιδεύσαμε διάφορους αλγόριθμους μηχανικής μάθησης με σκοπό την πρόβλεψη των θέσεων φωσφορυλίωσης του οργανισμού *Rattus norvegicus*. Τα περισσότερα υπολογιστικά μοντέλα που τρέξαμε σημειώνουν μια ακρίβεια κοντά στο 80%. Το μέγεθος της ακριβείας που επιτυγχάνουν οι αλγόριθμοι πρόβλεψης είναι αρκετά υψηλό αν ληφθεί υπόψη η σχετική ανεπάρκεια δεδομένων υψηλής ποιότητας. Με περισσότερα δεδομένα υψηλής ποιότητας ίσως οι αλγόριθμοι πρόβλεψης θα σημείωναν αρκετά μεγαλύτερη ακρίβεια. Το Keras/Tensorflow πέτυχε ακρίβεια της τάξης του 78.6%. Το Matlab πέτυχε ακρίβεια της τάξης του 79.8%. Από τους 7 αλγόριθμους του WEKA, η λογιστική παλινδρόμιση πέτυχε ακρίβεια της τάξης του 78.9%. Από τα παραπάνω συμπεραίνουμε ότι για το συγκεκριμένο πρόβλημα, οι διάφοροι αλγόριθμοι έχουν παρόμοια επιτυχία και ίσως ο πιο καθοριστικός παράγοντας είναι τα δεδομένα εκπαίδευσης και η ποιότητά τους. Επιπλέον, με βάση δύο άρθρα ανασκόπησης (39, 40) που μελετάνε δεκάδες διαφορετικά δημοσιευμένα εργαλεία πρόβλεψης θέσεων φωσφορυλίωσης που χρησιμοποιούν την πρωτοταγή δομή της ακολουθίας δεν είναι ειδικά για κινάσες, η ακρίβεια πρόβλεψης των καλύτερων αλγορίθμων γενικά κυμαίνεται γύρω στο 80%. Αυτό το επίπεδο ακρίβειας επιτεύχθηκε και στις δικές μας αναλύσεις. Η εργασία αυτή αποτελεί μέρος μιας συνολικότερης προσπάθειας του εργαστηρίου Βιοπληροφορικής του Τμήματος Βιοχημείας και Βιοτεχνολογίας να μελετήσει τις μετα-μεταφραστικές τροποποιήσεις και σε επόμενη φάση τα συμπεράσματα αυτής της εργασίας θα χρησιμοποιηθούν για την ανάπτυξη ενός web-server που θα προβλέπει θέσεις φωσφορυλίωσης στους ευκαρυώτες.

Βιβλιογραφία

1. Krüger, R., Kübler, D., Pallissé, R., Burkovski, A., Lehmann, W.D.: Protein and Proteome Phosphorylation Stoichiometry Analysis by Element Mass Spectrometry. Anal Chem. 78, 1987–1994 (2006).

2. Amoutzias, G.D., He, Y., Lilley, K.S., Van de Peer, Y., Oliver, S.G.: Evaluation and properties of the budding yeast phosphoproteome. Mol. Cell. Proteomics MCP. 11, M111.009555 (2012).

3. Cohen, P.: The origins of protein phosphorylation. Nat. Cell Biol. 4, E127–130 (2002).

4. Sadowski, I., Breitkreutz, B.-J., Stark, C., Su, T.-C., Dahabieh, M., Raithatha, S., Bernhard, W., Oughtred, R., Dolinski, K., Barreto, K., Tyers, M.: The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update. Database J. Biol. Databases Curation. 2013, bat026 (2013).

5. Amoutzias, G.D., Bornberg-Bauer, E., Oliver, S.G., Robertson, D.L.: Reduction/oxidationphosphorylation control of DNA binding in the bZIP dimerization network. BMC Genomics. 7, 107 (2006).

6. Papadopoulou, N., Chen, J., Randeva, H.S., Levine, M.A., Hillhouse, E.W., Grammatopoulos, D.K.: Protein kinase A-induced negative regulation of the corticotropin-releasing hormone R1alpha receptor-extracellularly regulated kinase signal transduction pathway: the critical role of Ser301 for signaling switch and selectivity. Mol. Endocrinol. Baltim. Md. 18, 624–639 (2004).

7. Oliveira, A.P., Sauer, U.: The importance of post-translational modifications in regulating Saccharomyces cerevisiae metabolism. FEMS Yeast Res. 12, 104–117 (2012).

8. Oliveira, A.P., Ludwig, C., Picotti, P., Kogadeeva, M., Aebersold, R., Sauer, U.: Regulation of yeast central metabolism by enzyme phosphorylation. Mol. Syst. Biol. 8, 623 (2012).

9. Deschênes-Simard, X., Kottakis, F., Meloche, S., Ferbeyre, G.: ERKs in cancer: friends or foes? Cancer Res. 74, 412–419 (2014).

10. Reimand, J., Wagih, O., Bader, G.D.: The mutational landscape of phosphorylation signaling in cancer. Sci. Rep. 3, 2651 (2013).

11. Lienhard, G.E.: Non-functional phosphorylations? Trends Biochem. Sci. 33, 351–352 (2008).

12. Landry, C.R., Levy, E.D., Michnick, S.W.: Weak functional constraints on phosphoproteomes. Trends Genet. TIG. 25, 193–197 (2009).

13. Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., Dunker, A.K.: The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 32, 1037– 1049 (2004).

14. Ingrell, C.R., Miller, M.L., Jensen, O.N., Blom, N.: NetPhosYeast: Prediction of Protein Phosphorylation Sites in Yeast. Bioinformatics. 23, 895–897 (2007).

15. Mok, J., Kim, P.M., Lam, H.Y.K., Piccirillo, S., Zhou, X., Jeschke, G.R., Sheridan, D.L., Parker, S.A., Desai, V., Jwa, M., Cameroni, E., Niu, H., Good, M., Remenyi, A., Ma, J.-L.N., Sheu, Y.-J., Sassi, H.E., 9 Sopko, R., Chan, C.S.M., De Virgilio, C., Hollingsworth, N.M., Lim, W.A., Stern, D.F., Stillman, B., Andrews, B.J., Gerstein, M.B., Snyder, M., Turk, B.E.: Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. Sci. Signal. 3, ra12 (2010).

16. Schwartz, D., Church, G.M.: Collection and motif-based prediction of phosphorylation sites in human viruses. Sci. Signal. 3, rs2 (2010).

17. Gauci, S., Helbig, A.O., Slijper, M., Krijgsveld, J., Heck, A.J.R., Mohammed, S.: Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. Anal. Chem. 81, 4493–4501 (2009).

18. Moser, K., and White, F.M. (2006). Phosphoproteomic analysis of rat liver by high capacity IMAC and LC-MS/MS. J. Proteome Res. *5*, 98–104.

19. Courcelles, M., Lemieux, S., Voisin, L., Meloche, S., and Thibault, P. (2011). ProteoConnections: a bioinformatics platform to facilitate proteome and phosphoproteome analyses. Proteomics *11*, 2654–2671.

20. Demirkan, G., Yu, K., Boylan, J.M., Salomon, A.R., and Gruppuso, P.A. (2011). Phosphoproteomic profiling of in vivo signaling in liver by the mammalian target of rapamycin complex 1 (mTORC1). PLoS ONE *6*, e21729.

21. Ferreira, R., Vitorino, R., Padrão, A.I., Espadas, G., Mancuso, F.M., Moreira-Gonçalves, D., Castro-Sousa, G., Henriques-Coelho, T., Oliveira, P.A., Barros, A.S., et al. (2014). Lifelong exercise training modulates cardiac mitochondrial phosphoproteome in rats. J. Proteome Res. *13*, 2045–2055.

22. Gündisch, S., Grundner-Culemann, K., Wolff, C., Schott, C., Reischauer, B., Machatti, M., Groelz, D., Schaab, C., Tebbe, A., and Becker, K.-F. (2013). Delayed times to tissue fixation result in unpredictable global phosphoproteome changes. J. Proteome Res. *12*, 4424–4434.

23. Han, D., Moon, S., Kim, Y., Ho, W.-K., Kim, K., Kang, Y., Jun, H., and Kim, Y. (2012a). Comprehensive phosphoproteome analysis of INS-1 pancreatic β-cells using various digestion strategies coupled with liquid chromatography-tandem mass spectrometry. J. Proteome Res. *11*, 2206–2223.

24. Han, D., Moon, S., Kim, Y., Ho, W.-K., Kim, K., Kang, Y., Jun, H., and Kim, Y. (2012b). Comprehensive phosphoproteome analysis of INS-1 pancreatic β-cells using various digestion strategies coupled with liquid chromatography-tandem mass spectrometry. J. Proteome Res. *11*, 2206–2223.

25. Hoffert, J.D., Wang, G., Pisitkun, T., Shen, R.-F., and Knepper, M.A. (2007). An automated platform for analysis of phosphoproteomic datasets: application to kidney collecting duct phosphoproteins. J. Proteome Res. *6*, 3501–3508.

26. Hou, C., Ma, J., Tao, D., Shan, Y., Liang, Z., Zhang, L., and Zhang, Y. (2010a). Organic-inorganic hybrid silica monolith based immobilized titanium ion affinity chromatography column for analysis of mitochondrial phosphoproteome. J. Proteome Res. *9*, 4093–4101.

27. Hou, J., Cui, Z., Xie, Z., Xue, P., Wu, P., Chen, X., Li, J., Cai, T., and Yang, F. (2010b). Phosphoproteome analysis of rat L6 myotubes using reversed-phase C18 prefractionation and titanium dioxide enrichment. J. Proteome Res. *9*, 777–788.

28. Lundby, A., Secher, A., Lage, K., Nordsborg, N.B., Dmytriyev, A., Lundby, C., and Olsen, J.V. (2012). Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. Nat Commun *3*, 876.

29. Michaelevski, I., Segal-Ruder, Y., Rozenbaum, M., Medzihradszky, K.F., Shalem, O., Coppola, G., Horn-Saban, S., Ben-Yaakov, K., Dagan, S.Y., Rishal, I., et al. (2010). Signaling to transcription networks in the neuronal retrograde injury response. Sci Signal *3*, ra53.

30. Nirujogi, R.S., Wright, J.D., Manda, S.S., Zhong, J., Na, C.H., Meyerhoff, J., Benton, B., Jabbour, R., Willis, K., Kim, M.-S., et al. (2015). Phosphoproteomic analysis reveals compensatory effects in the piriform cortex of VX nerve agent exposed rats. Proteomics *15*, 487–499.

31. Palmisano, G., Jensen, S.S., Le Bihan, M.-C., Lainé, J., McGuire, J.N., Pociot, F., and Larsen, M.R. (2012). Characterization of membrane-shed microvesicles from cytokine-stimulated β-cells using proteomics strategies. Mol. Cell Proteomics *11*, 230–243.

32. Shi, Z., Hou, J., Guo, X., Zhang, H., Yang, F., and Dai, J. (2013). Testicular phosphoproteome in perfluorododecanoic acid-exposed rats. Toxicol. Lett. *221*, 91–101.

33. Su, Z., Zhu, H., Zhang, M., Wang, L., He, H., Jiang, S., Hou, F.F., and Li, A. (2014). Salt-induced changes in cardiac phosphoproteome in a rat model of chronic renal failure. PLoS ONE *9*, e100331.

34. Tran, T., Park, J.-M., Kim, O.-H., Kim, B., Choi, D., Lee, J., Kim, K., Oh, B.-C., and Lee, H. (2013). Combined phospho- and glycoproteome enrichment in nephrocalcinosis tissues of phytate-fed rats. Rapid Commun. Mass Spectrom. *27*, 2767–2776.

35. Zhang, X., Højlund, K., Luo, M., Meyer, C., Geetha, T., and Yi, Z. (2012). Novel tyrosine phosphorylation sites in rat skeletal muscle revealed by phosphopeptide enrichment and HPLC-ESI-MS/MS. J Proteomics *75*, 4017–4026.

36. Chollet, F. (2018). Deep learning with Python (Shelter Island, New York: Manning Publications Co).

37.Hagan, M.T., Demuth, H.B., Beale, M.H., and De Jesus, O. Neural network design (Wrocław: Amazon Fulfillment Poland Sp. z o.o).

38.Witten, I.H., Frank, E., and Hall, M.A. (2011). Data mining: practical machine learning tools and techniques (Burlington, MA: Morgan Kaufmann).

39. Trost, B., and Kusalik, A. (2011). Computational prediction of eukaryotic phosphorylation sites. Bioinformatics *27*, 2927–2935.

40. Swaminathan,K. *et al.* (2010) Enhanced prediction of conformational flexibility and phosphorylation in proteins. *Adv. Exp. Med. Biol.*, **680**, 307–319.PP